情報処理学会論文誌　Vol. 49　No. 10　3694–3702 (Oct. 2008)

顔表情からの関心度推定に基づく映像コンテンツへのタギング

宮原正典[†1]　青木政樹[†1]
滝口哲也[†2]　有木康雄[†2]

近年，ユーザが視聴可能な映像コンテンツは莫大な量となってきているため，ユーザが自分の好きな映像コンテンツを探し出すことが困難になりつつある．そこで我々は，映像コンテンツを視聴するユーザを撮影し，その表情から関心度を推定することで映像コンテンツにタギングを行い，番組推薦に役立てるためのシステムを提案する．撮影された顔は，Elastic Bunch Graph Matching によって，顔特徴点抽出と個人認識が行われ，特定された個人に対して，Support Vector Machines によって関心のクラスが推定される．関心のクラスは，Neutral，Positive，Negative，Rejective の 4 種類であり，映像コンテンツと同期してフレームごとにタギングが行われる．評価実験の結果，関心クラス推定の平均再現率は 86.73% ，平均適合率は 86.67%となった．

Tagging Video Contents Based on Interest Estimation from Facial Expression

Masanori Miyahara,[†1] Masaki Aoki,[†1]
Tetsuya Takiguchi[†2] and Yasuo Ariki[†2]

Recently, there are so many videos available for people to choose to watch. To solve this problem, we propose a tagging system for video content based on facial expression that can be used for video content recommendations. Viewer's face captured by a camera is extracted by Elastic Bunch Graph Matching, and Interest class is estimated by Support Vector Machines. The interest classes are Neutral, Positive, Negative and Rejective. They are recorded as "interest tags" in synchronization with video content. Experimental results achieved an averaged recall rate of 86.73%, and averaged precision rate of 86.67%.

1. はじめに

近年，テレビでは多チャンネル化が進み，またインターネットでは，YouTube などに代表される動画共有サイトが発達してきたこともあり，ユーザが視聴できる映像コンテンツは莫大な量になっている．これにより，ユーザは，自分が見たい映像を自分で簡単に探し出すのが困難になりつつある．そこで，ユーザの好みに合わせて自動的に映像コンテンツを推薦してくれるシステムが期待されるが，そのためにはユーザがどういった映像コンテンツに関心があるのかを解析しなければならない．これに関する従来の研究には，大きく分けて 2 つの手法がある．1 つは，映像コンテンツそのものを解析する手法であり，もう 1 つは，映像コンテンツを視聴するユーザの振舞いを解析する手法である．

映像コンテンツそのものを解析する手法に関しては，映像に何が映っているか，映像がどのような意味を持っているか，ということが解析できれば，映像を推薦するうえで有用であると考えられる．映像解析の分野では，ショットの境界判定や，高次元特徴抽出，物体検出など様々な研究がなされているが，一般物体認識や，映像意図理解といったタスクは困難であるとされている[1)]．

一方，映像コンテンツを視聴するユーザの振舞いを解析する手法に関しては，好きなキーワードをユーザに登録してもらったり[2)]，リモコンの操作履歴を利用したりするといった手法が用いられる[3)]．しかし，これらの手法は，ユーザが自覚している好みしか反映できない．また定期的に好きなキーワードや俳優を登録するのはユーザにとって負担となる．

そこで，山本ら[4)] は，ユーザの映像視聴時の顔表情に着目し，HMM を用いて顔変化の時間的パターンに基づき，テレビ視聴者の興味区間を自動的に推定するシステムを提案した．これにより，ユーザに負担をかけずに，コンテンツのある区間に対する関心の有無を推定することが可能となった．

しかし，番組推薦を行うという視点に立った場合，関心がないという区間は，何も感じないという場合と，不快・嫌悪を感じる場合とが考えられ，これらを区別することによって推薦してほしくないシーンの判別に役に立つ．そのため，我々は，関心のクラス分類を，Neutral，Positive，Negative に拡張した．

†1 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University
†2 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University

さらに，より実環境に近づけると，ユーザは映像コンテンツの視聴中，ディスプレイを注視するだけでなく，違う方向を向いたり，口元を手で隠したりすることが考えられる．関心の推定には，正面を向いたときの顔特徴点の位置を用いているので，顔方向が変化したり，顔特徴点が隠れたりした場合，正しく推定できない．そのような推定結果が混入すると，番組推薦の精度が低下するため，除外したい．この点から，このような状況を自動的に識別し，Rejective というクラスに分類するようにした．

また，山本らは，顔表情の変化による特徴は，口の周囲に顕著に現れると考え，口の上下左右の特徴点の移動量に基づく特徴量を用いたが，Ekman ら[5] によると，目や眉の周囲にも重要な情報があることが知られている．そこで我々は，目や眉，唇の周囲の特徴点を高い精度で抽出するために，Elastic Bunch Graph Matching（EBGM）[6),7)] という手法を用いることにした．さらに，EBGM は顔特徴点抽出と個人認識を同じ枠組みで行えるという利点もある．

以上の点をふまえ，2 章で提案手法の概要について説明し，3 章，4 章，5 章で提案手法の詳細について述べる．6 章では実験結果とそれに対する考察を行い，7 章ではまとめを述べる．

2. 提案システムの概要

図 1 は，実験環境の上面図である．ユーザはディスプレイに映る映像コンテンツを 1 人で視聴している．ただし，本論文では，家庭内で 1 台のディスプレイを共同利用している環境を想定しているので，ユーザはつねに同一人物とは限らない．ユーザの顔はウェブカメラによって撮影される．PC は映像コンテンツの再生と，顔動画の解析処理を行う．

図 2 に，提案システムの流れを示す．まず初めに，Haar-like 特徴に基づく AdaBoost[8] によって，顔動画から顔領域を抽出する．これは，次の処理で計算時間を減らし，また顔領域のサイズを正規化するためである．次に，抽出された顔領域に対して，Gabor 特徴に基づく Elastic Bunch Graph Matching（EBGM）によって，顔特徴点の抽出と，個人認識を行う．そして，特定された個人に対して，あらかじめ作成しておいた個人顔表情モデルを選択する．個人顔表情モデルには，そのユーザの無表情顔特徴点と，個人顔表情識別器を登録しておく．このように，個人認識を行って，個人ごとに学習された顔表情識別器を用いる理由は，個人によって顔表情の表出の仕方が異なり，特に映像コンテンツ視聴時などの自然な表情の際には，それが顕著であるためである．最後に，ユーザの関心がどのクラスに属するかを Support Vector Machines（SVM）[9] によって推定する．SVM の特徴量には，無表情顔特徴点座標と，現在の表情の顔特徴点座標の差分を用いる．認識結果は「関心タグ」(Neutral, Positive, Negative, Rejective) として，映像コンテンツのフレームごとに同期して記録される．

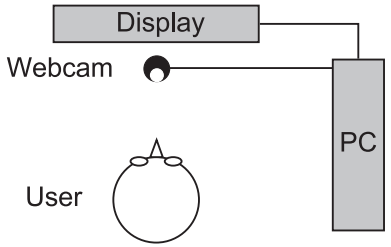

図 1　実験環境上面図
Fig. 1　Top view of experimental environment.

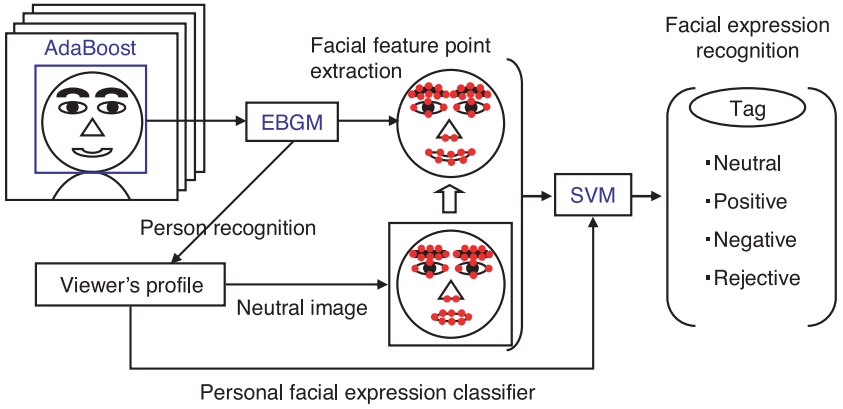

図 2　提案システムの流れ
Fig. 2　System flow.

3. AdaBoost による顔領域抽出

顔領域を抽出する手法には，HSV 色空間から肌色領域を抽出し，円形度などを用いて頭部領域を特定する手法などもあるが，本論文では，Viola，Jones らによって提案された，Haar-like 特徴に基づく AdaBoost を用いた．これは，Haar-like 特徴という濃淡特徴を特

徴量とする弱学習器をカスケード型に組み合わせて，強学習器を構成する手法である．高精度で顔領域を検出でき，かつ実時間で動作するため，広く用いられている．また，抽出した顔領域を一定サイズにリサイズすることで，カメラとユーザ間の距離を正規化した．さらに，抽出された領域に対してのみ，その後の処理を行うため，計算時間を短縮することができる．

4. EBGM による顔特徴点抽出と個人認識

4.1 Gabor Wavelet

Gabor Wavelet は Elastic Bunch Graph Matching の基礎であり，本節ではそのアルゴリズムについて述べる．

4.1.1 Gabor Wavelet

Gabor Wavelet は周波数を変化させることにより，全体的な特徴から局所的な特徴まで抽出することができる．また方向を変化させることで Wavelet の向きに対応した特徴を得ることができる．

Gabor Wavelet は式 (1) で与えられる．この Wavelet と顔特徴点の近傍領域を畳み込むことにより特徴量を得る．この関数には，単純な波の周波数と方位の両方を表す波ベクトル $\boldsymbol{k}_j$ と平滑化のためのガウス関数を含んでいる．

$$\psi_j(\boldsymbol{x}) = \frac{||\boldsymbol{k_j}||^2}{\sigma^2} \exp\left(-\frac{||\boldsymbol{k_j}||^2||\boldsymbol{x}||^2}{2\sigma^2}\right)\left[\exp\left(i\boldsymbol{k}_j\boldsymbol{x}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{1}$$

$$\boldsymbol{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos\varphi_\mu \\ k_\nu \sin\varphi_\mu \end{pmatrix} \tag{2}$$

ここで，$k_\nu = 2^{-\frac{\nu+2}{2}}\pi$，$\varphi_\mu = \mu\frac{\pi}{8}$ であり，$\nu = 0$，1，2，3，4 の 5 空間周波数，$\mu = 0$，1，2，3，4，5，6，7 の 8 方位の Wavelet が得られる．

4.1.2 Jet

Gabor Wavelet による畳み込みの特徴量を Jet と呼ぶ．これは特徴点の複素 Gabor Wavelet 係数の集合であり，これらを比較することにより特徴点位置の推定を行う．Jet のイメージを**図 3** に示す．

本論文では，8 方位，5 空間周波数を持った Real part と Imaginary part の計 80 の Wavelet を用いている．したがって，Jet は Real part と Imaginary part から生成された 40 の複素係数を持っている．Jet は式 (3) のように書くことができる．

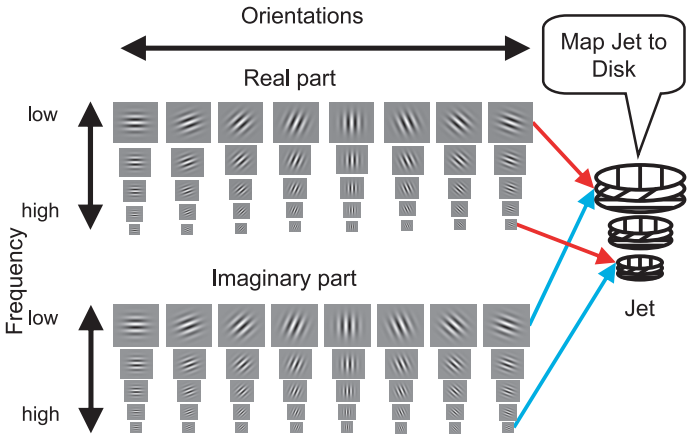

図 3　Gabor Wavelet と Jet
Fig. 3　Gabor Wavelet and Jet.

$$\mathcal{J}_j = a_j \exp(i\phi_j) \tag{3}$$

ここで，$\boldsymbol{x}$ は顔特徴点座標 (x, y) で $a_j(\boldsymbol{x})$ は複素係数の Magnitude，$\phi_j(\boldsymbol{x})$ は複素係数の Phase，j は 0〜39 の計 40 個である．

4.2 Jet の類似度

2 つの Jet，J と J' の相関を考える．2 つの Jet の位置は，画像上で $\boldsymbol{x}$ と $\boldsymbol{x}'$ であり，位置の差ベクトルは，

$$\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}' = \begin{pmatrix} dx \\ dy \end{pmatrix} \tag{4}$$

である．

ここで，2 つの Jet の大きさと位相に関する類似度を考える．すなわち，

$$S_D(J, J') = \frac{\sum_{j=0}^{N-1} a_j a'_j \cos(\phi_j - \phi'_j)}{\sqrt{\sum_{j=0}^{N-1} a_j^2 \sum_{j=0}^{N-1} a_j'^2}} \tag{5}$$

この式を用い J の Jet に最も類似した J' を見つけるには，$\phi_j - \phi'_j = \boldsymbol{k}_j\boldsymbol{x} - \boldsymbol{k}_j\boldsymbol{x}' = \boldsymbol{k}_j(\boldsymbol{x} - \boldsymbol{x}') = \boldsymbol{k}_j\boldsymbol{d}$ であることに注目し，位相差 $\phi_j - \phi'_j - \boldsymbol{k}_j\boldsymbol{d}$ を考慮に入れると，

図 4　顔特徴点
Fig. 4　Facial feature points.

$$S_D(J,J') = \frac{\sum_{j=0}^{N-1} a_j a'_j \cos(\phi_j - \phi'_j - \boldsymbol{k_j d})}{\sqrt{\sum_{j=0}^{N-1} a_j^2 \sum_{j=0}^{N-1} a_j'^2}} \tag{6}$$

となり，すべての j に対して位相だけでなく，大きさも含めた類似度が最大になる $\boldsymbol{d}$ を求めればよい．

次に，この $\boldsymbol{d}$ の推定手法について述べる．まず初期探索点の場所の Jet の類似度を，$dx = dy = 0$ として式 (6) で計算し，その後，初期探索点の右上，右下，左下，左上の点における Jet の類似度と初期探索点の Jet の類似度を比較し，類似度が高いほうを次の探索開始点として，さらに右上，右下，左下，左上と探索を続けていく．また探索点が周りの点よりも高い類似度の場合はステップ幅を 1/2 にしてさらに同じことを繰り返し，局所的な探索に入る．これを収束するまで繰り返し，その点を特徴点位置として抽出する．

4.3　EBGM

本節では，Elastic Bunch Graph Matching について述べる．

4.3.1　Graph

本論文では，顔特徴点は，**図 4** で示されるような，表情の特徴として重要とされる[5] 眉，目，唇の周囲を主とする 34 点を用いることにした．Graph とは，この 34 点の特徴点位置から Jet を抜き出したもののことを意味する．Graph のイメージを**図 5** に示す．この抽出された顔領域における Graph を利用することで，目的領域の形状把握が可能になる．

4.3.2　Bunch Graph

Bunch Graph とは複数人から Graph を抽出し，それを束（Bunch）状にした Graph のことである．この Bunch Graph をもとに，対象の特徴点位置の探索を行う．Bunch Graph は束にした Graph の情報を保持し，探索のときに束にしてあるものと探索点での Jet を比較したのち，束の中から類似度の一番高い Jet を採用して位置決めをする．

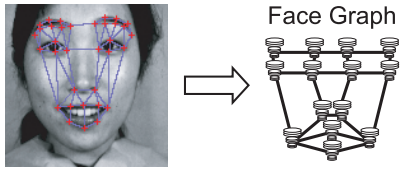

図 5　顔特徴点から抽出された Face Graph
Fig. 5　Face Graph extracted from facial feature points.

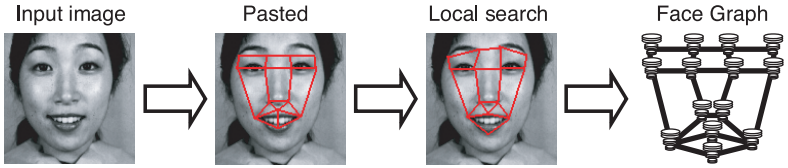

図 6　EBGM の流れ
Fig. 6　Elastic Bunch Graph Matching procedure.

4.3.3　Elastic Bunch Graph Matching

事前に手動で顔特徴点を指定して生成された Bunch Graph は，特徴点位置探索に使われる．Elastic Bunch Graph Matching の流れ図を**図 6** に示す．まず始めに，特徴点を検出したい画像を入力すると，そこに Bunch Graph が貼り付き，そこから 4.2 節で示した手法を用いて特徴点位置の局所探索に入る．位置の推定が完了した後，入力された顔画像に対する Face Graph が抽出され，顔特徴点座標が得られる．

また個人認識は，あらかじめ登録されている Face Graph G と，入力された顔画像から得られた Face Graph G' の Jet の類似度の総和を式 (7) で計算し，この値が一番高いものを本人として決定する．M は顔特徴点の数で，$M = 34$ である．

$$S_{jet}(G,G') = \frac{1}{M}\sum_{j=0}^{M-1} S_D(\mathcal{J}_j, \mathcal{J}'_j) \tag{7}$$

4.4　スムージング

前項の手法により，入力された顔画像に対して個人認識を行うことができるようになったが，本論文では，視聴者の顔を撮影した動画を入力し，毎フレーム個人認識を行う．さらに，現フレームと，前後それぞれ W フレームの，合計 $2W+1$ フレームにおいて，個人認識結果のスムージングを行い，現フレームの視聴者を特定する．すなわち，n 番目のフレームに

おける，視聴者 p に対する類似度 S_{p_n} に対して，スムージング後の類似度を S'_{p_n} とすると，

$$S'_{p_n} = \frac{\sum_{i=n-W}^{n+W} S_{p_i}}{2W+1} \tag{8}$$

であり，n 番目のフレームでの個人認識結果は $\arg\max_p(S'_{p_n})$ となる．

これは，視聴者がごくわずかな時間で交代を繰り返すことがないという前提に基づき，個人認識結果の瞬間的な誤りを除去するために行うものである．なお，個人認識には，顔特徴点抽出に用いた Face Graph を利用するため，システム全体の計算量には大きな影響は与えない．

5. SVM による関心クラスの推定

5.1 関心クラスの定義

従来では，関心の有無のみを推定していたので，関心のクラスは，「関心あり」「関心なし」と分類されていた．しかし，「関心なし」には，その映像に対して，特に何も感じていない場合と，嫌悪や不快を感じてもう見たくない場合があると考えられる．特に，インターネットにおける動画共有サイトでは，多くの一般人が動画をアップロードして，プロによる編集作業を経てないものが多いため，テレビとは違って様々な映像コンテンツが存在する．そこで，**表 1** のような関心クラスを定義した．これにより，協調フィルタリング[10] のような手法を用いることで，見たい番組のみを推薦し，見たくない番組を推薦しない，といったシステムを構築することが期待される．また，実環境では，ユーザはつねにディスプレイを注視しているとは限らず，顔を傾けたり，口元に手をやったりして，顔特徴点がつねに見えているとは限らない．そのため，そのような状態を表す，Rejective というクラスを定義した．Rejective と判断された区間は，関心推定の信頼性が低いと判断し，番組推薦を行うときには除外する．

5.2 Support Vector Machines

Support Vector Machines（SVM）は，学習データを，特徴空間上において，マージン最大化基準で分類する超平面を構成し，識別境界とする手法である．本論文では，RBF カーネルによってカーネル化し，one-against-the-rest 法によって多クラスの識別を実現した．

5.3 特徴ベクトル

SVM の特徴ベクトルには，無表情のときと表情があるときの顔特徴点の移動量を用いた．すなわち，あらかじめ個人ごとに登録しておいた無表情画像における顔特徴点座標を

表 1　関心クラスの定義
Table 1　Interest classes.

Classes	Meanings
Neutral (Neu)	Expressionless
Positive (Pos)	Happiness, Laughter, Pleasure, etc.
Negative (Neg)	Anger, Disgust, Displeasure, etc.
Rejective (Rej)	Not watching the display in the front direction, occluding part of face, tilting the face, etc.

(x_{Ai}, y_{Ai}) とし，各フレームの表情画像における顔特徴点座標を (x_{Bi}, y_{Bi}) とすると，特徴ベクトル $\boldsymbol{m}$ は以下のように定義される．

$$\boldsymbol{m_i} = [x_{Bi} - x_{Ai} \;\; y_{Bi} - y_{Ai}]^T \tag{9}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \;\; \boldsymbol{m_2} \;\; \cdots \;\; \boldsymbol{m_{34}}]^T \tag{10}$$

5.4 Rejective の推定

Rejective に関しても他のクラスと同様に，学習データを用意して，SVM で推定を行う．Rejective なフレームは，顔特徴点抽出に失敗することが多いと考えられるが，そのような学習データを集めることで，失敗しているかどうかが SVM で推定できるという考え方である．ただし，前段の処理である AdaBoost による顔領域抽出で，顔が検出されなかったフレームは，無条件で Rejective と推定する．

5.5 スムージング

関心クラスはフレームごとに推定しているが，人間の表情はフレーム単位で高速に変化することは少ないため，4.4 節と同様に式 (8) と類似の式を用いることで，瞬間的な認識誤りの除去を行う．本論文では，人間の表情の持続時間を考慮し，現フレームと前後それぞれ $W = 7$ フレーム，すなわち 1 秒の区間でスムージングを行うことにする．

6. 実　　験

6.1 実験条件

本論文では，家庭内で 1 台のディスプレイを共同利用し，それぞれ別の時間帯に映像コンテンツを視聴しているという環境を想定している．そのため，一般的な世帯人数[11] であると考えられる，3 名の被験者 A，B，C に対して，図 1 に示すような実験環境で，4 本の映像コンテンツを見せた．被験者には，顔表情を誇張したり，抑制したりしないように指示した．映像コンテンツは，30 分のテレビ番組から CM などを取り除き，本編だけを抽出した

図 7　タギング用のインタフェース
Fig. 7　Tagging interface.

表 2　タギング結果
Table 2　Tagged results (frames).

	Neu	Pos	Neg	Rej	Total
Subject A	49,865	7,665	3,719	1,466	62,715
Subject B	56,531	2,347	3,105	775	62,758
Subject C	56,271	4,653	155	1,748	62,827

ものを利用し，1 本あたり約 17 分間の長さである．映像コンテンツのジャンルは「バラエティ番組」を用いた．これは，「ドラマ」や「ニュース」などと比べ，視聴者に表情変化が頻繁に見られるからである．被験者が映像コンテンツを見ている間，システムは，映像コンテンツと同期して，毎秒 15 フレームでユーザの顔を録画し続ける．その後，被験者は，自分の顔動画と映像コンテンツの両方を見ながら，表 1 に従って，映像コンテンツにタギングを手動で行う．映像コンテンツにタギングを行う際，**図 7** のようなインタフェースを用いた．左のウィンドウには映像コンテンツとユーザの顔動画が再生されていて，右のウィンドウにはタギング用のボタンがある．このようなインタフェースを用いて手動タギングを行うことで，被験者が，どのような表情をしたときに，どのような関心を持ったかという正解ラベルをシステムに与えることができる．このラベルは，顔表情に基づく関心度推定において，学習と評価に用いる．被験者による手動タギングの結果を，**表 2** に示す．この手動タグ付きの顔動画を実験動画と呼び，以降の評価実験に用いる．

6.2　AdaBoost による顔領域抽出

実験動画のすべてのフレームに対して，Haar-like 特徴に基づく AdaBoost によって顔領域の抽出実験を行った．顔が検出されなかったフレームの数は，**表 3** のとおりである．

これらのフレームは，すべて手動で Rejective とタギングされたフレームであり，顔が傾き，カメラに正対していないフレームが多かった．これは，正面の顔画像のみを用いて AdaBoost の学習を行っているからだと考えられる．

表 3　顔領域が未検出のフレーム数
Table 3　No facial region extraction (frames).

	No extraction
Subject A	378
Subject B	248
Subject C	637

表 4　顔領域抽出の実験結果（被験者 A）
Table 4　Facial region extraction experiment for subject A.

	Neu	Pos	Neg
False extraction	20	3	1
Total frames	49,865	7,665	3,719
Rate (%)	0.040	0.039	0.027

表 5　顔領域抽出の実験結果（被験者 B）
Table 5　Facial region extraction experiment for subject B.

	Neu	Pos	Neg
False extraction	132	106	9
Total frames	56,531	2,347	3,105
Rate (%)	0.234	4.516	0.290

表 6　顔領域抽出の実験結果（被験者 C）
Table 6　Facial region extraction experiment for subject C.

	Neu	Pos	Neg
False extraction	11	0	1
Total frames	56,271	4,653	155
Rate (%)	0.020	0.000	0.645

また，顔が検出されたフレームに関しては，それが正しい顔領域かどうかを人間が目で見て確認した．実験結果を**表 4**，**表 5**，**表 6** に示す．

表が示すように，Neutral，Positive，Negative に対する平均誤検出率は，被験者 A で 0.0354%，被験者 B で 1.68%，被験者 C で 0.222%となった．被験者 B の Positive のときの誤検出率が高いのは，被験者 B は笑ったときに大きく顔を動かす癖があり，そのため顔が検出しにくくなったものと考えられる．

表 7　個人認識の実験結果
Table 7　Person recognition experiment.

	Subject A	Subject B	Subject C	Sum	Recall (%)
Subject A	62,230	67	40	62,337	99.83
Subject B	28	62,476	6	62,510	99.94
Subject C	378	3,262	58,550	62,190	94.15
Sum	62,636	65,805	58,597	187,038	
Precision (%)	99.35	94.94	99.92		

6.3　EBGM による個人認識

AdaBoost によって正しく検出された全フレームに対して，EBGM によって個人認識を行った．実験結果を**表 7** に示す．ただし，この結果は，4.4 節で述べたスムージング処理を行う前のものである．

なお，表中の再現率（Recall）と適合率（Precision）は，以下で定義される．

$$Recall = \frac{Relevant\ frames \cap Estimated\ frames}{Relevant\ frames} \tag{11}$$

$$Precision = \frac{Relevant\ frames \cap Estimated\ frames}{Estimated\ frames} \tag{12}$$

一部の被験者で再現率，適合率が低下した理由は，登録画像に対して，角度や顔表情が大きく変化した結果，他人の顔と類似度が高くなってしまったからだと考えられる．

ただし，この個人認識結果に対して，現フレームと前後それぞれ $W = 75$ フレーム，すなわち約 10 秒間の区間でスムージングを行ったところ，再現率と適合率はすべての被験者において 100%となった．

この実験により，提案システムはユーザの個人表情モデルを正しく選択できることが確認できた．

6.4　SVM による関心クラスの推定

実験動画のすべてのフレームに対して，SVM を用いて関心クラスの推定を行った．4 本の実験動画のうち 3 本を学習データに用いて，残りの 1 本をテストデータとして，クロスバリデーションを行った．

6.4.1　評 価 尺 度

評価尺度としては，前節で定義した再現率（Recall）と適合率（Precision）を用いた．

また，実験動画は Neutral のフレームが圧倒的に多いので，フレーム単位で平均を求めると，平均値は Neutral の結果にほぼ依存してしまう．そのため，クラス単位で平均を求めることにする．すなわち，平均再現率（averaged recall）と，平均適合率（averaged precision）は以下のように表される．

$$Averaged\ recall = \frac{\Sigma(Neu, Pos, Neg, Rej)recall}{4} \tag{13}$$

$$Averaged\ precision = \frac{\Sigma(Neu, Pos, Neg, Rej)precision}{4} \tag{14}$$

表 8　関心クラス推定の実験結果（被験者 A）
Table 8　Confusion matrix for subject A.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	48,275	443	525	622	49,865	96.81
Pos	743	6,907	1	14	7,665	90.11
Neg	356	107	3,250	6	3,719	87.39
Rej	135	0	5	1,326	1,466	90.45
Sum	49,509	7,457	3,781	1,968	62,715	
Precision (%)	97.51	92.62	85.96	67.38		

表 9　関心クラス推定の実験結果（被験者 B）
Table 9　Confusion matrix for subject B.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	56,068	138	264	61	56,531	99.18
Pos	231	2,076	8	32	2,347	88.45
Neg	641	24	2,402	38	3,105	77.36
Rej	203	0	21	551	775	71.10
Sum	57,143	2,238	2,695	682	62,758	
Precision (%)	98.12	92.76	89.13	80.79		

6.4.2　関心クラス推定の実験結果

実験結果を，**表 8**，**表 9**，**表 10** に示す．ただし，Rejective の結果は，SVM で Rejective と推定されたフレーム数と，AdaBoost で顔が検出されなかったフレーム数の合計である．

全被験者に対する平均再現率は 86.73%，平均適合率は 86.67%となった．被験者の表情表出が小さい場合，被験者が Positive，Negative とタギングしていても，Neutral と誤認識されることが多かった．また，今回は 1 つのフレームに，Neutral，Positive，Negative，Rejective という 4 つの関心クラスのうちいずれか 1 つしか存在しないと仮定したため，中間的な表情をしていた場合，誤認識が多く発生していた．また，同じ映像コンテンツを見せても，関心の持ち方は個人によってばらつきがあった．そのため，特定の関心クラスが極端

表 10　関心クラス推定の実験結果（被験者 C）
Table 10　Confusion matrix for subject C.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	55,030	866	27	348	56,271	97.79
Pos	318	4,331	0	4	4,653	93.08
Neg	50	4	101	0	155	65.16
Rej	262	9	10	1,467	1,748	83.92
Sum	55,660	5,210	138	1,819	62,827	
Precision (%)	98.87	83.13	73.19	80.65		

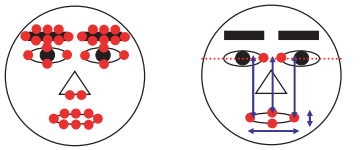

図 8　提案手法と従来手法の特徴量
Fig. 8　Proposal and conventional feature.

表 11　従来手法の特徴量を用いた実験結果（被験者 A）
Table 11　Confusion matrix for subject A by conventional feature.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	49,344	306	82	133	49,865	98.96
Pos	832	6,815	0	18	7,665	88.91
Neg	3,055	494	154	16	3,719	4.141
Rej	462	104	0	900	1,466	61.39
Sum	53,693	7,719	236	1,067	62,715	
Precision (%)	91.90	88.29	65.25	84.35		

に少ないと，学習データが十分に集まらず，誤認識が多くなると考えられる．

6.4.3　特徴量の比較実験結果

さらに，特徴量として，従来手法[4] で用いられている，**図 8** の右の顔中の矢印で示される 5 次元の特徴ベクトルを用いて，それ以外は前項と同じ条件で比較実験を行った．実験結果を，**表 11**，**表 12**，**表 13** に示す．

全被験者に対する平均再現率は 59.16%，平均適合率は 83.41%となり，提案手法の 86.73%，86.67%を下回った．特に，Negative のときの性能低下が著しい．これは，笑顔や喜びなど

表 12　従来手法の特徴量を用いた実験結果（被験者 B）
Table 12　Confusion matrix for subject B by conventional feature.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	56,358	55	44	74	56,531	99.69
Pos	518	1,748	71	10	2,347	74.48
Neg	1,937	231	937	0	3,105	30.18
Rej	415	49	2	309	775	39.87
Sum	59,228	2,083	1,054	393	62,758	
Precision (%)	95.15	83.92	88.90	78.63		

表 13　従来手法の特徴量を用いた実験結果（被験者 C）
Table 13　Confusion matrix for subject C by conventional feature.

	Neu	Pos	Neg	Rej	Sum	Recall (%)
Neu	55,387	403	1	480	56,271	98.43
Pos	2,952	1,691	1	9	4,653	36.34
Neg	107	1	8	39	155	5.16
Rej	463	19	0	1,266	1,748	72.43
Sum	58,909	2,114	10	1,794	62,827	
Precision (%)	94.02	79.99	80.00	70.57		

の表情では，口の周囲の情報だけでも十分な特徴を含んでいるが，嫌悪や不快などの表情では，目や眉の周囲に重要な特徴が現れるため，と考えられる．

7. おわりに

本論文では，ユーザの顔表情からの関心度推定に基づき，Neutral，Positive，Negative というタグを映像コンテンツに付与するシステムを提案した．加えて，関心度の推定が困難なフレームに対して，学習された SVM によって，自動的に Rejective というタギングを行った．関心クラスの推定について，3 名の被験者による評価実験を行った結果，平均再現率と平均適合率は約 87%となった．これにより，提案システムが，映像を見たユーザの興味のある区間にタギングを行い，番組推薦を行う際に有用な情報を提供できることが示された．今後は，より多くの被験者や，バラエティ番組以外の映像を用いて実験を行っていく必要がある．その際，システム利用前の手動タギングの負担を減らすために，システムが，類似した表情をクラスタリングよってひとまとめにしてからユーザに提示する，といった手法を考えている．さらに，Neutral，Positive，Negative，Rejective 以外のもっと詳細な表情も認識できるようにする予定である．また，現在，音声や視聴動作など，顔表情以外のマ

ルチモーダル情報も組み合わせて，実際に番組推薦を行うシステムの構築を検討中である．

参 考 文 献

1) Smeaton, A.F., Over, P. and Kraaij, W.: Evaluation campaigns and TRECVid, *Proc. 8th ACM International Workshop on Multimedia Information Retrieval* (*MIR '06*), Santa Barbara, California, USA, Oct. 26–27, 2006, pp.321–330, ACM Press, New York, NY (2006).
2) 隆　朋也，渡辺　尚，樽口秀昭：履歴情報を用いたTV番組選択支援エージェント，情報処理学会論文誌，Vol.42, No.12, pp.3130–3143 (2001).
3) 益満　健，越後富夫：映像重要度を用いたパーソナライズ要約映像作成手法，電子情報通信学会論文誌，Vol.J84-D-II, No.8, pp.1848–1855 (2001).
4) 山本　誠，谷本浩昭，新田直子，馬場口登：個人的選好獲得のための特定人物のテレビ視聴時における興味区間推定，電子情報通信学会論文誌，Vol.J90-D-II, No.8, pp.2202–2211 (2007).
5) Ekman, P. and Friesen, W.V.（著），工藤　力（訳編）：表情分析入門，誠信書房 (1987).
6) Wiskott, L., Fellous, J.-M., Kruger, N. and Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.775–779 (1997).
7) Bolme, D.S.: Elastic Bunch Graph Matchin, *Partial fulfillment of the requirements for the Degree of Master of Science Colorado State University Fort Collins*, Colorado (Summer 2003).
8) Viola, P. and Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Kauai*, USA, pp.1–9 (2001).
9) Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer, Heidelberg (1995).
10) Resnick, P., et al.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. CSCW '94*, pp.175–186 (1994).
11) 総務省統計局：社会・人口統計体系．http://www.stat.go.jp/data/ssds/index.htm
12) Michael, J.L., Akamatsu, S., Kamachi, M. and Gyoba, J.: Coding Facial Expressions with Gabor Wavelets, *Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society, pp.200–205 (1998).

(平成20年2月19日受付)
(平成20年7月 1 日採録)

宮原　正典

2007年神戸大学工学部情報知能工学科卒業．現在，同大学院工学研究科情報知能学専攻博士課程前期課程在学中．顔表情認識の研究に従事．電子情報通信学会学生会員．

青木　政樹

2007年神戸大学工学部情報知能工学科卒業．現在，同大学院工学研究科情報知能学専攻博士課程前期課程在学中．顔認識の研究に従事．

滝口　哲也（正会員）

1999年奈良先端科学技術大学院大学博士後期課程修了．1999年日本アイ・ビー・エム東京基礎研究所．2004年神戸大学工学部講師．博士（工学）．音情報処理，画像処理等の研究に従事．日本音響学会，電子情報通信学会，IEEE各会員．

有木　康雄（正会員）

1974年京都大学工学部情報工学科卒業．1976年同大学院修士課程修了．1979年同大学院博士課程修了．1980年京都大学工学部情報工学科助手．1990年龍谷大学理工学部電子情報学科助教授，1992年同教授．2003年神戸大学工学部教授．工学博士．1987〜1990年エディンバラ大学客員研究員．画像処理，音声情報処理に従事．電子情報通信学会，映像情報メディア学会，日本音響学会，人工知能学会，画像電子学会，IEEE各会員．