C1-15

顔表情クラスタリングによる映像コンテンツへのタギング
Tagging for Video Contents Based on User's Facial Expression Clustering

宮原 正典 † **青木 政樹** † **滝口 哲也** ‡ **有木 康雄** ‡
Masanori Miyahara Masaki Aoki Tetsuya Takiguchi Yasuo Ariki

1. はじめに

近年，テレビの多チャンネル化や，インターネットにおける動画共有サイトの発達により，ユーザが視聴できる映像コンテンツの量は増加し続けている．これにより，ユーザは，自分が見たい番組を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 [1] が必要となってきている．我々は，以前，ユーザの顔表情に基づいて，映像コンテンツにタギングを行うシステム [2] を提案したが，このシステムは事前に顔表情を個人ごとに学習させる必要があった．そこで，今回，顔表情を自動的にクラスタリングし，ユーザにとって少ない負担で利用できるシステムを提案する．

2. システムの概要

本研究では，まず，図 1 に示すような，PC のディスプレイの映像をユーザが 1 人で視聴している実験環境を構築した．ウェブカメラはユーザの顔を撮影していて，PC は映像の再生とユーザの顔動画の解析処理を行う．

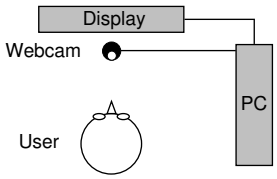

図 1: Top view of experimental environment

ウェブカメラで撮影された顔動画は，まず，顔動画から Haar-like 特徴を用いた AdaBoost 法 [3] によって，正確な顔領域を切り出す．そして切り出された顔領域に対して Gabor 特徴を用いた Elastic Bunch Graph Matching(EBGM)[4] により，顔特徴点座標を抽出する．そして，あらかじめ登録されたユーザの無表情顔画像から抽出された特徴点座標と，毎フレーム抽出される特徴点座標の差分を特徴量とする．この特徴量系列を全てクラスタリングし，各クラスタにユーザは 1 度だけタグを与えることで，映像コンテンツの全フレームに，顔表情に基づくタギングを行うことができる．

3. システムの手法

3.1 特徴量抽出

特徴量抽出の流れを図 2 に示す．

†神戸大学大学院工学研究科
‡神戸大学自然科学系先端融合研究環

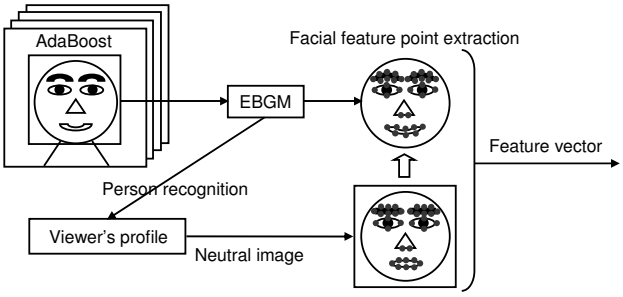

図 2: Feature vector extraction

EBGM は，図 3 のように，まず，画像を様々な周波数と方向を持った Gabor フィルターで畳み込み，それらの応答の集合を Jet とし，Bunch Graph と呼ばれるモデルを作成する．この Bunch Graph は，複数人のデータから作成しておく．次に，Bunch Graph と，入力画像の各特徴点の Jet との間で類似度を計算し，特徴点の座標を推定する．本研究では，Bunch Graph の特徴点は，図 2 中に示すような 34 点を用いた．

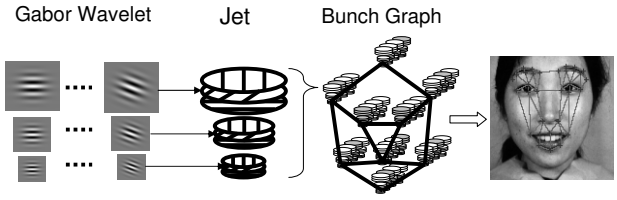

図 3: Elastic Bunch Graph Matching

そして，EBGM によって抽出された 34 点の顔特徴点座標と，あらかじめ保持しておいた本人の無表情画像から抽出した顔特徴点座標の差分を取ることにより，68 次元の顔特徴点移動量ベクトルを求め，これを特徴量とした．

3.2 顔表情クラスタリング

テレビ視聴時の顔表情は，意図的に演じられた顔表情ではないので，自然であり，表出強度がそれほど強くなく，さらに個人によって表出の仕方に差があると考えられる．そのため，複数人から学習された表情認識器を用いると，各個人に対しては，認識性能が低下する．そこで，システム利用者は事前に，映像コンテンツを視聴して，それに正解のタグをつけて学習データとする，という作業が必要であった．しかし，自然な表情はいつ表出されるかわからないため，ある程度長時間の視聴が必要であり，それを逐一チェックして正解のタグをつける作業はユーザにとって負担となっていた．

そこで，今回提案するシステムは，映像コンテンツを視聴するユーザを一定時間撮り続けて，その顔表情をクラスタリングし，いくつかの顔表情クラスタを形成してユーザに提示するものである．ユーザはフレームではなく，その顔表情クラスタに正解のタグをつければよい．顔表情クラスタリングのイメージを表 4 に示す．

本研究では，顔表情クラスを隠れトピック，顔特徴点座標を単語，そのフレームでの特徴量を文書，と考えて，pLSA[5] によって，あるフレームでの顔表情を，隠れトピックの発生確率で表現する．

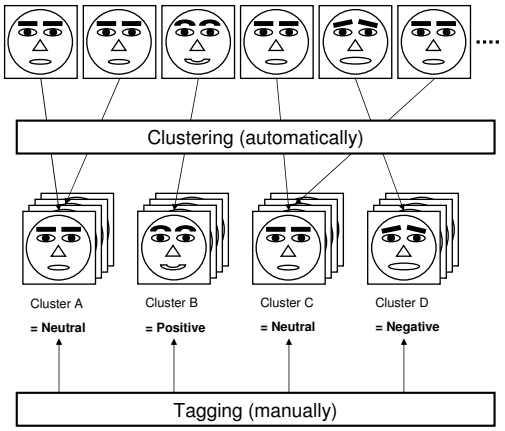

図 4: Facial expression clustering

4. 評価実験

4.1 実験条件

図 1 に示すような実験環境を用いて，被験者 1 名に 1 回約 17 分間の映像コンテンツを 4 回分視聴させた．この際，被験者の顔動画を映像コンテンツと同期させながら，毎秒 15 フレームで記録した．その後，被験者に自分の顔動画と映像コンテンツを並べて見せ，表 1 に基づく手動タギングを毎フレーム行わさせた．これを正解ラベルとする．

表 1: Facial expression classification

クラス	内容
Neutral(Neu)	無表情
Positive(Pos)	喜び，笑い，感動，など
Negative(Neg)	怒り，嫌悪，不快，など
Reject(Rej)	画面に顔を向けていない，顔の一部が隠れている，顔が傾いている，など

一方，顔動画の全フレームを用いて，隠れトピック数を 10 として，pLSA を行い，各フレームでの隠れトピックの混合比率を求めた．ただし，本実験では，もっとも発生確率の高いトピックを，そのフレームの属するクラスタとした．トピックの混合を認めない理由は，被験者が各顔表情の混合比率をタギングして正解ラベルを作成するのが困難だからである．その後，各クラスタに分類された顔画像を表示し，表 1 から最も近いと思う表情を各クラスタに 1 つ選択させた．このクラスタに属する全てのフレームはその表情にタギングされたと考え，手動で与えた正解ラベルがどれだけ一致するかを調べた．

4.2 実験結果

実験結果を表 2 に示す．

表 2: Confusion matrix

	Neu	Pos	Neg	Rej	Sum	Recall
Neu	47362	1459	449	595	49865	94.98
Pos	1018	6585	48	14	7665	85.91
Neg	575	590	2496	58	3719	67.11
Rej	206	133	80	1047	1466	71.42
Sum	49161	8767	3073	1714	62715	
Precision	96.34	75.11	81.22	61.09		

上表より，各表情の平均適合率 78.44%，平均再現率は 79.86%となった．

4.3 考察

全フレームに対して手動でタギングを行ったときと比較して，顔表情クラスタリングを行い，各クラスタにタギングを行った場合，タギングの性能は約 79%になるが，タギングの手間は大幅に減少する．

5. おわりに

本研究では，映像コンテンツにタギングを行う際に，コンテンツを視聴したユーザの顔表情を，あらかじめクラスタリングを行ってからユーザに提示することを提案した．これにより，より多種の表情をタギングしようとしたときに，ユーザの負担を大幅に低減させることが可能となる．今後は，被験者の数や表情の種類を増やした実験を行う予定である．

参考文献

[1] 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).

[2] Masanori Miyahara, Masaki Aoki, Tetsuya Takiguchi, Yasuo Ariki: “ Tagging Video Contents with Positive/Negative Interest Based on User ’s Facial Expression ”, The 14th International Multimedia Modeling Conference (MMM2008), pp.210-219, 2008-01.

[3] Viola, P. and Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Kauai, USA, pp.1-9 (2001).

[4] Wiskott, L., Fellous, J.-M., Kruger, N. and Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), pp.775-779 (1997).

[5] T. Hofmann: “ Probabilistic Latent Semantic Indexing, ” SIGIR, pp.50-57, 1999.