映像視聴時の顔器官の動き情報を用いた関心の度合いの推定に関する検討

斉藤 直輝*，山本 浩太郎*，西川 大地**，原川 良介**，岩橋 政宏**，浅水 仁*

A Note on Estimation of Interest-level to Videos Using Facial Expression

Naoki SAITO, Kotaro YAMAMOTO, Daichi NISHIKAWA, Ryosuke HARAKAWA, Masahiro IWAHASHI, Satoshi ASAMIZU

Abstract －This paper presents an estimation method of interest-level to videos using facial expression. The proposed method combines visual and facial expression features via Local Discriminative Canonical Correlation Analysis (LDCCA). LDCCA can derive suitable features for interest-level estimation by considering the class information and data structure. Consequently, the proposed method realizes accurate estimation of interest-level to videos. The experimental results show the performance of the proposed method and shows its discussion.

Key words: interest-level estimation, facial expression, local discriminative canonical correlation analysis.

１．はじめに

近年，情報通信技術の発達と移動体ブロードバンド環境の普及により，YouTube[1]やNETFLIX[2]などの映像配信サービスの利用者が急激に増加している[1]．これらの利用者は，サービス内で配信されている映像を視聴することができる．しかしながら，映像配信サービスは，大量かつ多様な映像を配信しているため，利用者が関心のある映像を見つけ出し，視聴することは容易ではない．この問題を解決するために，利用者にとって関心がある映像を自動で推薦する手法の実現が期待されている．

映像配信サービスにおける映像推薦を実現するために，利用者の映像に対する関心の度合いを推定する手法が種々提案されている[2,3]．これらの手法では，映像から算出された特徴量 (映像特徴量) に加え，映像視聴時の脳波や視線などの生体情報に基づく特徴量を利用して関心の度合いを推定している．生体情報は，映像に対する関心の度合いとの関連性が高いため，映像特徴量と協調的に用いることで，推定精度が向上している．ここで，従来手法では，関心の度合いを推定する際に，推定対象となる映像と生体情報が必要となる．したがって，利用者が視聴したことがない映像に対する関心の度合いを予め推定し，映像推薦に利用することは困難である．

上記の問題を解決するため，以前我々は映像視聴時の顔器官の動き情報から算出された特徴量 (顔器官特徴量) を用いた関心の度合いの推定手法を提案した[4].

* 釧路工業高等専門学校
** 長岡技術科学大学

[1] https://www.youtube.com/
[2] https://www.netflix.com/

具体的には，正準相関分析 (Canonical Correlation Analysis: CCA) [5]により映像特徴量を顔器官特徴量との相関が最大となる空間へ射影することで得られた正準映像特徴量を利用して，関心の度合いを推定した．CCAによる射影行列を事前に算出することで，映像に対する関心の度合いの推定に顔器官特徴量を必要としないため，利用者が視聴したことがない未知の映像に対する関心の度合いを推定することが可能になる．しかしながら，CCAは正準映像特徴量を算出する際，関心の度合いに関する情報を利用していないため，射影により得られた特徴量が，推定に適しているとは限らない．さらに，顔器官の動き情報に含まれている外れ値の影響により，関心の度合いの推定精度が低下するため，外れ値に頑健な特徴量の算出方法の導入が必要とされている．

そこで本稿では，Local Discriminative CCA (LDCCA)[6] により映像および顔器官特徴量を同時に用いた関心の度合いの推定手法を提案する．提案手法では，推定に顔器官特徴量を利用することで，映像特徴量のみでは捉えられない情報を推定に用いることが可能となる．また，LDCCAにより映像特徴量を顔器官特徴量との相関が最大となる新変量空間へ関心の度合いに関するクラス情報を考慮しながら射影することで，弁別性が高く，外れ値に頑健な推定が可能となる．

以降では，2章で提案手法について説明し，3章で提案手法の有効性を検証するために行った実験の結果について説明する．最後に4章でまとめとする．

２．映像に対する関心の度合いの推定

本章では，提案手法について説明する．提案手法ではまず，映像および映像を視聴しているユーザの顔器官の動き情報から映像特徴量および顔器官特徴量をそれぞれ算出する．さらに，LDCCA により，映像特徴量を顔器官特徴量との相関が最大となる新たな特徴空間へ射影するための射影行列を導出する．提案手法では，新変量空間へ射影することで算出される正準映像特徴量を用いて映像に対する関心の度合いを推定する．以降では，2.1 節で映像および顔器官特徴量の算出方法について説明し，2.2 節で LDCCA による射影行列の算出方法について説明する．さらに，2.3 節で正準映像特徴量の算出方法について説明する．

2.1. 映像および顔器官特徴量の算出

本節では，映像および顔器官特徴量の算出方法について説明する．提案手法では，ImageNet[7]を用いて学習された畳み込みニューラルネットワーク[8]であるInception-v3 [9]の第 3 プーリング層から出力される2,048 次元のベクトルをフレーム毎に算出し，これらの平均ベクトルを映像特徴量とする．

次に，顔器官特徴の算出方法を説明する．提案手法ではまず，文献[10]に基づいて，映像視聴時のユーザの顔の映像から 68 点の特徴点を検出する．これらの特徴点について，映像を視聴している間の縦および横方向の動きの標準偏差を並べた 126 次元のベクトルを顔器官特徴量とする．

2.2. LDCCA による射影行列の算出

本節では，LDCCA による射影行列の算出方法について説明する． 今，i $(= 1, 2, \dots, N; N$は総映像数$)$ 番目の映像について算出された映像特徴量および顔器官特徴量をそれぞれ$\boldsymbol{x}_i \in \mathbb{R}^{2048}$および$\boldsymbol{y}_i \in \mathbb{R}^{126}$とする．また，これらの特徴量を並べた行列を$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{2048\times N}$および$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N] \in \mathbb{R}^{126\times N}$とする．$\boldsymbol{X}$および$\boldsymbol{Y}$が事前に中心化されているものとするとき，LDCCA では，クラス内の相関を最大化し，クラス間の相関を最小化する新変量空間への射影ベクトル$\widehat{\boldsymbol{w}}_x$および$\widehat{\boldsymbol{w}}_y$を次式により算出する．

$$\left\{\widehat{\boldsymbol{w}}_x, \widehat{\boldsymbol{w}}_y\right\} = \arg\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \boldsymbol{w}_x^{\mathrm{T}} \tilde{\boldsymbol{C}}_{xy} \boldsymbol{w}_y \quad (1)$$

$$\text{s.t.}\ \boldsymbol{w}_x^{\mathrm{T}} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \boldsymbol{w}_x = \boldsymbol{w}_y^{\mathrm{T}} \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{w}_y = 1$$

ただし，$\tilde{\boldsymbol{C}}_{xy}$は以下の通りに定義される．

$$\tilde{\boldsymbol{C}}_{xy} = \boldsymbol{C}_w - \eta \boldsymbol{C}_b \quad (2)$$

ηはクラス内およびクラス間の局所的な相関を表す行列$\boldsymbol{C}_w$および$\boldsymbol{C}_b$の重みを調整するパラメータである．なお，$\boldsymbol{C}_w$および$\boldsymbol{C}_b$は以下の通りに定義される．

$$\boldsymbol{C}_w = \sum_{i=1}^{N} \sum_{\boldsymbol{x}_k \in \mathcal{N}_x^I(\boldsymbol{x}_i), \boldsymbol{y}_k \in \mathcal{N}_y^I(\boldsymbol{y}_i)} \boldsymbol{x}_i \boldsymbol{y}_k^{\mathrm{T}} + \boldsymbol{x}_k \boldsymbol{y}_i^{\mathrm{T}} \quad (3)$$

$$\boldsymbol{C}_b = \sum_{i=1}^{N} \sum_{\boldsymbol{x}_k \in \mathcal{N}_x^E(\boldsymbol{x}_i), \boldsymbol{y}_k \in \mathcal{N}_y^E(\boldsymbol{y}_i)} \boldsymbol{x}_i \boldsymbol{y}_k^{\mathrm{T}} + \boldsymbol{x}_k \boldsymbol{y}_i^{\mathrm{T}} \quad (4)$$

ただし，$\mathcal{N}_x^I(\boldsymbol{x}_i)$および$\mathcal{N}_y^I(\boldsymbol{y}_i)$はそれぞれ$i$番目の映像と同一クラスに属し，$\boldsymbol{x}_i$および$\boldsymbol{y}_i$の$K$近傍に存在する映像に対する特徴量の集合を表す．さらに，$\mathcal{N}_x^E(\boldsymbol{x}_i)$お

よび$\mathcal{N}_y^E(\boldsymbol{y}_i)$はそれぞれ$i$番目の映像とは異なるクラスに属し，$\boldsymbol{x}_i$および$\boldsymbol{y}_i$の$K$近傍に存在するサンプルの特徴量の集合を表す．式(1)の制約付き最適化問題は，ラグランジュの未定乗数法を用いて解くことができる．具体的には，次式の通りに定義されるラグランジュ関数$L(\boldsymbol{w}_x, \boldsymbol{w}_y, \lambda_x, \lambda_y)$の$\boldsymbol{w}_x$, $\boldsymbol{w}_y$, λ_x, およびλ_yについての極値問題を解く．

$$L(\boldsymbol{w}_x, \boldsymbol{w}_y, \lambda_x, \lambda_y) = \boldsymbol{w}_x^{\mathrm{T}} \tilde{\boldsymbol{C}}_{xy} \boldsymbol{w}_y + \lambda_x (1 - \boldsymbol{w}_x^{\mathrm{T}} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \boldsymbol{w}_x) + \lambda_y (1 - \boldsymbol{w}_y^{\mathrm{T}} \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{w}_y) \quad (5)$$

式(5)のラグランジュ関数を$\boldsymbol{w}_x$および$\boldsymbol{w}_y$で微分して 0 とおくと，以下の通りに表される．

$$\tilde{\boldsymbol{C}}_{xy} \boldsymbol{w}_y = 2\lambda_x \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \boldsymbol{w}_x \quad (6)$$

$$\tilde{\boldsymbol{C}}_{xy}^{\mathrm{T}} \boldsymbol{w}_x = 2\lambda_y \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{w}_y \quad (7)$$

式(6)および(7)をまとめると，式(1)の制約付き最適化問題は，最終的に次式の一般化固有値問題に帰着される．

$$\begin{bmatrix} \boldsymbol{0} & \tilde{\boldsymbol{C}}_{xy} \\ \tilde{\boldsymbol{C}}_{xy}^{\mathrm{T}} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} \quad (8)$$

式(8)を解くことで得られる固有値に対応する固有ベクトル$\widehat{\boldsymbol{w}}_x^d$ ($d \in \{1, 2, \ldots, D\}$; Dは固有値問題を解くことで得られた固有ベクトルの数) を並べることで射影行列$\widehat{\boldsymbol{W}} = [\widehat{\boldsymbol{w}}_x^1, \widehat{\boldsymbol{w}}_x^2, \cdots, \widehat{\boldsymbol{w}}_x^D] \in \mathbb{R}^{2048 \times D}$を算出する．

2.3. 正準映像特徴量の算出

本節では，前節で算出した射影行列を用いた正準映像特徴量の算出方法および関心の度合いの推定方法について説明する．提案手法では，射影行列$\widehat{\boldsymbol{W}}$を用いて，次式で定義される射影により正準映像特徴量$\hat{\boldsymbol{x}}_i$を算出する．

$$\hat{\boldsymbol{x}}_i = \widehat{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{x}_i \quad (9)$$

新たに得られた映像より算出された特徴量$\boldsymbol{x}_{\mathrm{test}}$に対しても式(9)と同様に射影を行うことで，新たに顔器官の動き情報を取得することなく正準映像特徴量$\hat{\boldsymbol{x}}_{\mathrm{test}}$を算出できる．提案手法では正準映像特徴量と映像に対する関心の度合いを用いることで，教師あり学習により推定器を構築し，映像に対する関心の度合いの推定を行う．

3．実験

本章では，提案手法の有効性を検証するために行った実験の結果について説明する．本実験では，YouTube より無作為に取得した 15 秒間の広告映像 112 本を用いた．これらの映像を用いて 5 名の実験参加者から映像視聴時の顔器官の動き情報を取得した．各参加者は，映像を 1 本ずつ視聴し，広告映像で紹介されていた商品を“4: 購入したい”，“3: やや購入したい”，“2: やや購入したくない”，“1: 購入したくない”の 4 段階の評価を行った．本実験では，評価値が 4 もしくは 3 の映像を「関心あり」，2 もしくは 1 の映像を「関心なし」として，関心の有無の識別を行った．本実験では関心の有無の識別方法として，Support Vector Machine (SVM) [11]を用いた．ただし，SVM のカーネル関数は線形カーネルとした．

表 1. 提案手法と比較手法の概要.

	映像特徴量	顔器官特徴量	特徴統合方法
提案手法	○	○	LDCCA
比較手法1	○		-
比較手法2		○	-
比較手法3	○	○	**ベクトル結合**
比較手法4	○	○	CCA
比較手法5	○	○	DCCA

本実験では，leave-one-out 交差検証を行い，評価指標として，次式に示す F 値を用いた．

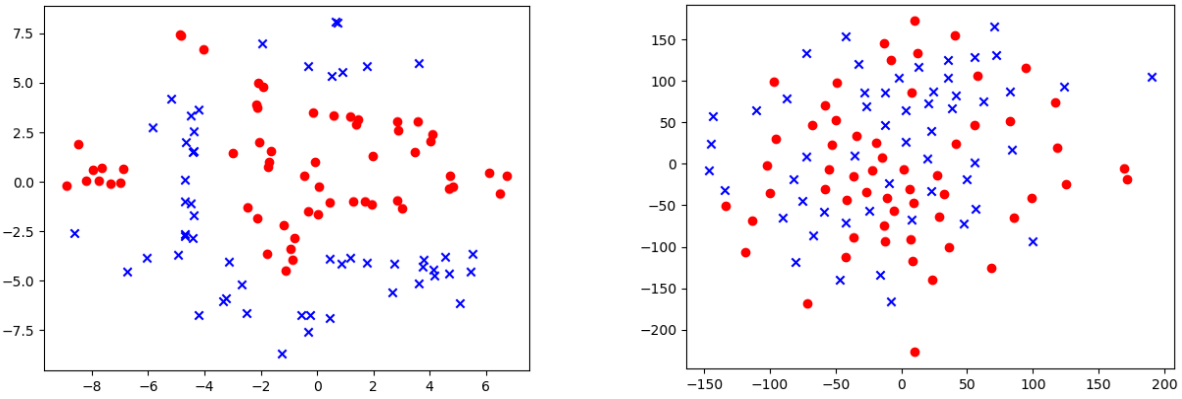
$$\text{F値} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (10)$$

ただし，TP，FP，および FN はそれぞれ True Positive, False Positive，および False Negative のサンプル数を表す．また，提案手法の有効性を検証するため，表 1 に示す 5 種類の比較手法を用いた．具体的には，比較手法 1 および 2 より，映像特徴量と顔器官特徴量を同時に用いることの有効性を確認する．また，比較手法 3 および 4 より，提案手法において LDCCA を導入することの有効性を確認する．さらに，クラス情報を考慮できる CCA である Discriminative CCA (DCCA) [12]を用いた比較手法 5 より，LDCCA により特徴量空間の局所性を考慮して射影を算出する有効性を確認する．なお，DCCA には射影後の特徴量の次元数がクラス数と等しくなるという制約が存在する．本実験における，正準映像特徴量の次元数等の CCA に関するパラメータは，識別精度が最も高くなるように実験的に設定した．

提案手法および比較手法による実験参加者毎の識別結果の F 値を表 2 に示す．実験結果より，すべての実験参加者において，提案手法による識別結果の F 値が

表 2. SVM による広告映像に対する関心の有無の識別結果.

	提案手法	比較手法1	比較手法2	比較手法3	比較手法4	比較手法5
参加者1	**0.98**	0.79	0.82	0.83	0.86	0.89
参加者2	**0.86**	0.52	0.65	0.54	0.57	0.68
参加者3	**0.89**	0.78	0.84	0.80	0.80	0.84
参加者4	**0.99**	0.86	0.91	0.85	0.89	0.91
参加者5	**0.82**	0.62	0.58	0.59	0.66	0.76
平均	**0.91**	0.71	0.76	0.72	0.76	0.82



(a) LDCCA により射影された特徴量　　(b) CCA により射影された特徴量

図 2. LDCCA および CCA により射影された正準映像特徴量の可視化結果．赤丸印および青バツ印はそれぞれ「関心あり」および「関心なし」のサンプルを表す．なお，正準映像特徴量は，t-SNE により 2 次元へ圧縮している．

他の手法を上回っていることが確認できる．以降，識別結果に基づいて，提案手法と各比較手法との比較を行い，提案手法の有効性について考察を行っていく．

まず，比較手法 1 および 2 よりも提案手法の F 値の平均が大きく上回っていることから，映像特徴量および顔器官特徴量を同時に用いて推定することの有効性が確認できる．さらに特徴量をベクトル結合により統合する比較手法 3 よりも CCA を利用した比較手法 4 の F 値が高いことから, CCA に基づく方法により新たな特徴量空間への射影を求めることで，異なる種類の特徴量を統合する有効性が確認できる．また，提案手法の F 値の平均が比較手法 3 および 4 よりも上回っていることから，提案手法に LDCCA を導入することの有効性が確認できる．ここで，図 2 に LDCCA および CCA により射影された正準画像特徴量を t-SNE[13]により 2 次元へ圧縮してプロットした散布図を示す．LDCCA により射影された特徴量は, CCA と比較して，「関心あり」と「関心なし」のサンプルが分離されていることが確認できる．この理由として，LDCCA ではクラス情報を考慮していることが考えられる．したがって，LDCCA により射影された正準画像特徴量は，従来の CCA と比較して，弁別性が高いため関心の有無の識別精度も向上したと考えられる．最後に，提案手法の F 値の平均が比較手法 5 を上回っていることから，LDCCA により特徴量空間の局所性を考慮して射影を求める有効性が確認できる．射影を求める際に局所性を考慮することで，外れ値に頑健な特徴量を識別に利用することが可能となるため，識別精度が向上したと考えられる．以上より，提案手法の有効性が確認された．

4．まとめ

本稿では，LDCCA により映像特徴量および顔器官特徴量を同時に用いた映像に対する関心の度合いの推定手法を提案した．また，実際に映像と顔器官の動き情報を用いた実験結果より，従来の CCA を利用した手法と比較して, 推定精度の向上を確認した. 今後は,

新たな特徴量の算出に multimodal Similarity Gaussian Process latent variable model (m-SimGP) [14]を導入することで，推定精度の向上を目指す.

謝辞

本研究の一部は，令和 2 年度高専-長岡技科大共同研究の助成により行われた.

参考文献

[1] 総務省, “令和元年度版 情報通信白書,” p. 49, 2019.

[2] S. H. Fairclough, A. J. Karran, and K. Gilleade,“Classification accuracy from the perspective of the user: real-time interaction with physiologicalcomputing,” in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3029-3038.

[3] S. Liu, J. Lv, Y. Hou, T. Shoemaker, Q. Dong, K. Li, and T. Liu, “What makes a good movie trailer?: interpretation from simultaneous eeg and eyetracker recording,” in *Proceedings of the 24th ACM international conference on multimedia*, 2016, pp. 82-86.

[4] K. Yamamoto, D. Nishikawa, N. Saito, R. Harakawa, M. Iwahashi, and S. Asamizu, “Interest level estimation to video using facial expression-aware visual features,” in *Proceedings of the 5th STI-Gigaku*, 2020, p. 65.

[5] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[6] Y. Peng, D. Zhang, and J. Zhang, “A new canonical correlation analysis algorithm with local discrimination,” *Neural processing letters*, vol. 31, no. 1, pp. 1-15, 2010.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[10] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867-1874.

[11] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[12] T. Sun, S. Chen, J. Yang, and P. Shi, “A novel method of combined feature extraction for recognition,” in *Proceedings of the 8th IEEE international conference on data mining*, 2008, pp. 1043-1048.

[13] L. Maaten, and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. 11, pp. 2579-2605, 2008.

[14] G. Song, S. Wang, Q. Huang, Q. Tian, “Multimodal similarity gaussian process latent variable model,” *IEEE Transactions on image processing*, vol. 26, no. 9, pp. 4168-4181, 2017.