LETTER

Human-Centered Video Feature Selection via mRMR-SCMMCCA for Preference Extraction

Takahiro OGAWA[†a)], ***Member***, **Yoshiaki YAMAGUCHI**[†], ***Nonmember***, **Satoshi ASAMIZU**[††], ***and*** **Miki HASEYAMA**[†], ***Members***

SUMMARY This paper presents human-centered video feature selection via mRMR-SCMMCCA (minimum Redundancy and Maximum Relevance-Specific Correlation Maximization Multiset Canonical Correlation Analysis) algorithm for preference extraction. The proposed method derives SCMMCCA, which simultaneously maximizes two kinds of correlations, correlation between video features and users' viewing behavior features and correlation between video features and their corresponding rating scores. By monitoring the derived correlations, the selection of the optimal video features that represent users' individual preference becomes feasible.

key words: *Canonical Correlation Analysis, feature selection, preference extraction, viewing behavior*

1. Introduction

In recent years, extraction of users' individual preference has become necessary for realizing successful videos retrieval and recommendation [1]. In general, even if different users provide the same rating information (rating scores) for the same videos, their individual preference for these videos may be different since each video contains several objects. Since users' viewing behavior such as gazing, facial expression and body movements indicates the users' attention, it becomes one of the most important factors to extract the users' individual preference. Thus, there have been proposed several methods predicting rating scores of videos on the basis of the users' viewing behavior [2]. Although video features that are closely related to each user's attention are different from each other, no previously reported methods consider this point.

It is necessary to perform selection of video features which can reflect each user's individual preference. The study of feature selection has been intensively carried out, and many benchmarking and state-of-the-art methods such as mRMR (minimum Redundancy Maximum Relevance) algorithm [3] and mRMR-CCA (minimum Redundancy Maximum Relevance-Canonical Correlation Analysis) algorithm [4] have been proposed. Unfortunately, the above feature selection algorithms only monitor the relationship between two modalities, e.g., video features and their corresponding rating scores which represent preference degrees of videos. Although we have proposed feature selection algorithm using SLPCCA-OC (Supervised Locality Preserving Canonical Correlation Analysis with Ordinal Classes) algorithm [5], it is difficult to simultaneously use correlations between video features and "viewing behavior features and rating scores of users".

In this paper, we present an mRMR-SCMMCCA (minimum Redundancy and Maximum Relevance-Specific Correlation Maximization Multiset Canonical Correlation Analysis) feature selection algorithm for video preference extraction. SCMMCCA simultaneously maximizes the two kinds of correlations centered at the video features to achieve the feature selection for extracting the individual video preference. Specifically, we try to find the best video feature set by solving an optimization problem maximizing "relevance represented by the maximized correlations" and minimizing "redundancy represented by the correlation among features". This feature selection algorithm using SCMMCCA is called "mRMR-SCMMCCA algorithm", and it is the biggest contribution of this paper. Consequently, the users' individual preference can be extracted as the selection of the optimal video features by using this non-conventional algorithm.

2. Video Feature Selection via mRMR-SCMMCCA

This section shows the mRMR-SCMMCCA feature selection algorithm. First, we explain extraction of video and viewing behavior features in Sect. 2.1. Furthermore, the specific feature selection algorithm, which is the biggest contribution of this paper, is presented in Sect. 2.2.

2.1 Feature Extraction from Video and Viewing Behavior

From a training dataset, the proposed method calculates video features $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, N)$ and their corresponding user's viewing behavior features $\boldsymbol{y}_i$, where $N = \sum_{i=1}^{M} n_i$, and M is the number of training videos and n_i is the number of frames in ith video, i.e., N becomes the number of all training samples. Due to the limitation of spaces, we only show the overview of the calculation of these features below.

Video features (1209 dimensions):
As shown in Table 1, we adopt 145 audio features consist-

Manuscript received June 19, 2016.
Manuscript revised October 10, 2016.
Manuscript publicized November 4, 2016.

[†]The authors are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
[††]The author is with National Institute of Technology, Kushiro College, Kushiro-shi, 084–0916 Japan.
a) E-mail: ogawa@lmd.ist.hokudai.ac.jp

DOI: 10.1587/transinf.2016EDL8126

Table 1 Video features and viewing behavior features used in the proposed method.

	Feature quantities	Dimensions
Video features	Audio features consisting of Dynamics, Spectral, Timbre, Tonal and Rhythm obtained by [6]	145
	HSV color histogram	64
	Bag of visual words based on SURF [7]	1000
Viewing behavior features (face)	2D rectangle region of the face	2
	3D angle of the face	3
	3D movement of the head position	3
	Facial expression descriptor based on Action unit [8]	6
Viewing behavior features (body movement)	Distance between the user's centroid and a display	1
	2D movement of the user's centroid	2
	2D rectangle region of the body	2
	3D movement of both hands' position	6
	3D movement of both legs' position	6
	Angle of the body based on distance between both shoulders and a display	3

ing of Dynamics, Spectral, Timbre, Tonal and Rhythm obtained by MIRToolbox which is used for music feature extraction [6]. Furthermore, HSV color histogram and Bag of visual words based on SURF [7] are calculated as the visual features. Then the video feature vector $\boldsymbol{x}_i \in \mathbb{R}^{D_x}$ is obtained from each ith sample, where $D_x = 1209$ from Table 1.

Viewing Behavior Features (34 dimensions):
We obtain facial features and body movement features as shown in Table 1 by using a Kinect sensor. To calculate the facial features, we detect landmark points on the face, and a 3D face model[†] corresponding to the landmark points is automatically extracted by the Kinect. Then we can obtain the 14-dimensional facial features. For calculating the body movement features, we obtain a user's region and coordinates of the user's skeleton from the Kinect. Then 20-dimensional body movement features are calculated as shown in Table 1. Finally, we obtain the viewing behavior feature vector $\boldsymbol{y}_i \in \mathbb{R}^{D_y}$ for each ith sample, where $D_y = 34$ from Table 1.

2.2 Derivation of mRMR-SCMMCCA Algorithm

This section presents the mRMR-SCMMCCA feature selection algorithm that is derived on the basis of the relationship estimation between video features and "viewing behavior features and rating scores". First, we define a video feature matrix $\boldsymbol{X} \in \mathbb{R}^{D_x \times N}$ and a viewing behavior feature matrix $\boldsymbol{Y} \in \mathbb{R}^{D_y \times N}$ as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N]$, respectively. Furthermore, the corresponding user's rating scores representing degrees of video preference are defined as $l_i \in \{1, 2, \cdots, R\}$ $(i = 1, 2, \cdots, N)$, and $\boldsymbol{l} = [l_1, l_2, \cdots, l_N] \in \mathbb{R}^{1 \times N}$ is also defined, where R is the number of grades. In SCMMCCA, we try to solve the following optimization problem, which maximizes the sum of the two kinds of correlations, the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (video,behavior) and the correlation between $\boldsymbol{X}$ and $\boldsymbol{l}$ (video,rating), to obtain the optimal projections $\boldsymbol{w}_x \in \mathbb{R}^{D_x}$, $\boldsymbol{w}_y \in \mathbb{R}^{D_y}$ and $w_l \in \mathbb{R}^1$:

$$\begin{aligned} &\arg \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, w_l} \boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{Y}^T \boldsymbol{w}_y + \boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{l}^T w_l \\ &\quad \text{s.t. } \boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w}_x + \boldsymbol{w}_y^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}_y + w_l \boldsymbol{l}\boldsymbol{l}^T w_l = 1. \end{aligned} \quad (1)$$

As shown in the above equation, SCMMCCA tries to maximize the sum of the two kinds of correlations of (video,behavior) and (video,rating). Whereas multiset CCA tries to maximize the sum of the correlations of all pairs [9], SCMMCCA maximizes the sum of specific correlations, i.e., the sum of the two kinds of correlations of (video,behavior) and (video,rating). Note that if we do not use the vector $\boldsymbol{l}$, we can monitor only the relationship between the video features and the viewing behavior features. Since the viewing behavior is caused from visual and audio stimuli, we have to distinguish the viewing behavior features related to the preference and those not related to the preference. Therefore, in order to remove the features unrelated to the user's preference, we need to use the vector $\boldsymbol{l}$.

From the Lagrange multiplier approach, the optimal projections are obtained by solving the following generalized eigenvalue problem:

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{X}\boldsymbol{Y}^T & \boldsymbol{X}\boldsymbol{l}^T \\ \boldsymbol{Y}\boldsymbol{X}^T & \boldsymbol{0} & 0 \\ \boldsymbol{l}\boldsymbol{X}^T & \boldsymbol{0} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \\ w_l \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{X}\boldsymbol{X}^T & \boldsymbol{0} & 0 \\ \boldsymbol{0} & \boldsymbol{Y}\boldsymbol{Y}^T & 0 \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{l}\boldsymbol{l}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \\ w_l \end{bmatrix}. \quad (2)$$

The eigenvalues $\lambda_d^{\text{SCMMCCA}}$ $(d = 1, 2, \cdots, D_x + D_y + 1;$ $\lambda_d^{\text{SCMMCCA}} > \lambda_{d+1}^{\text{SCMMCCA}})$ corresponding to λ which are obtained by solving the above eigenvalue problem represent the strength of the relationship between the video features and "the viewing behavior features and the rating scores". The proposed method defines

$$\rho_{\text{SCMMCCA}}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{l}) = \lambda_1^{\text{SCMMCCA}} \quad (3)$$

as a new criterion corresponding to the relevance between

[†]This 3D face model outputs head poses and facial expression descriptor based on Action Units [8], and the Microsoft Face Tracking Software Development Kit for Kinect for Windows (Face Tracking SDK) supports six Action Units (Upper lip raiser, Jaw lowerer, Lip stretcher, Brow lowerer, Lip corner depressor, and Outer brow raiser).

Table 2 Quantitative evaluation of the proposed method and the previously reported methods.

	Our method		mRMR-CCA (video,rating)		mRMR-CCA (video,behavior)		SLPCCA-OC		mRMR	
	MAE	MZE	MAE	MZE	MAE	MZE	MAE	MZE	MAE	MZE
Subject1	**0.771**	**0.625**	0.822	0.637	0.956	0.644	0.952	0.642	0.956	0.644
Subject2	0.719	0.557	**0.672**	**0.547**	0.788	0.621	0.814	0.611	0.750	0.593
Subject3	**0.949**	**0.692**	0.994	0.706	1.323	0.776	1.134	0.784	1.175	0.754
Subject4	**0.896**	**0.653**	0.913	0.664	1.088	0.699	1.041	0.683	1.067	0.700
Subject5	**0.553**	**0.511**	0.575	0.527	0.622	0.600	0.620	0.593	0.622	0.600
Average	**0.778**	**0.608**	0.795	0.616	0.955	0.668	0.912	0.663	0.914	0.658

the video features and the two user-related features for selecting the optimal video features.

By using the relevance criterion defined in Eq. (3), we derive a new feature selection algorithm, i.e., the mRMR-SCMMCCA algorithm. Specifically, we perform the optimal feature selection one-by-one in the same manner as the previously reported feature selection algorithms [3], [4]. Specifically, the selection of kth optimal video feature is performed as

$$\max_{\boldsymbol{m}_{d_x}\in\Omega_X-S_{k-1}}\left[\rho_{\mathrm{SCMMCCA}}(\boldsymbol{m}_{d_x},\boldsymbol{Y},\boldsymbol{l})-\rho_{\mathrm{CCA}}(\boldsymbol{m}_{d_x},\hat{\boldsymbol{S}}_{k-1})\right], \tag{4}$$

where $\boldsymbol{m}_{d_x}\in\mathbb{R}^{1\times N}$ $(d_x=1,2,\cdots,D_x)$ is a vector including d_xth row of $\boldsymbol{X}$. Furthermore, Ω_X is a set of all video features $\boldsymbol{m}_{d_x}$ $(d_x=1,2,\cdots,D_x)$, S_{k-1} is a set of video features selected in the previous $k-1$ iterations, and $\hat{\boldsymbol{S}}_{k-1}\in\mathbb{R}^{(k-1)\times N}$ is a matrix whose rows are these $k-1$ selected features. Note that $\rho_{\mathrm{SCMMCCA}}(\boldsymbol{m}_{d_x},\boldsymbol{Y},\boldsymbol{l})$ and $\rho_{\mathrm{CCA}}(\boldsymbol{m}_{d_x},\hat{\boldsymbol{S}}_{k-1})$ respectively correspond to the relevance and the redundancy of the features.

In Eq. (4), $\rho_{\mathrm{CCA}}(\cdot,\cdot)$ is a function which outputs canonical correlation between two multi-dimensional variates. Given two arbitrary matrices $\boldsymbol{A}\in\mathbb{R}^{D_A\times N}$ and $\boldsymbol{B}\in\mathbb{R}^{D_B\times N}$, the specific definition of $\rho_{\mathrm{CCA}}(\boldsymbol{A},\boldsymbol{B})$ is given as

$$\rho_{\mathrm{CCA}}(\boldsymbol{A},\boldsymbol{B})=\max_{\boldsymbol{w}_a,\boldsymbol{w}_b}\frac{\boldsymbol{w}_a^T\boldsymbol{A}\boldsymbol{B}^T\boldsymbol{w}_b}{\sqrt{\boldsymbol{w}_a^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{w}_a}\sqrt{\boldsymbol{w}_b^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{w}_b}}, \tag{5}$$

where $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$ are projections maximizing the canonical correlation between the two variates $\boldsymbol{A}$ and $\boldsymbol{B}$. By using the Lagrange multiplier approach, we obtain the optimal solutions of $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$ and their corresponding correlation coefficients λ_d^{CCA} $(d=1,2,\cdots,\min(D_A,D_B);\lambda_d^{\mathrm{CCA}}>\lambda_{d+1}^{\mathrm{CCA}})$. Then $\rho_{\mathrm{CCA}}(\boldsymbol{A},\boldsymbol{B})$ in Eq. (5) becomes λ_1^{CCA}. As shown in Eq. (5), it corresponds to the solution of the general CCA problem.

As shown in Eq. (4), the mRMR-SCMMCCA algorithm can find the optimal video features which have the maximum correlation with the two user-related features and the minimum correlation each other.

Although the metric space used for $\rho_{\mathrm{SCMMCA}}(\boldsymbol{m}_{d_x},\boldsymbol{Y},\boldsymbol{l})$ and $\rho_{\mathrm{SCMMCA}}(\boldsymbol{m}_{\tilde{d}_x},\boldsymbol{Y},\boldsymbol{l})$ $(d_x\neq\tilde{d}_x)$ can be different, we try to monitor the maximum correlation existing between the three variables. The aim of the proposed method is to find the video features which have the biggest relationship with the viewing behavior and the rating scores. Therefore, we simply monitor the maximum correlation obtained by performing the SCMMCCA as shown in Eq. (3). This idea was also adopted in the mRMR-CCA algorithm [4].

3. Experimental Results

In order to verify the effectiveness of our method, this section shows experimental results. In this experiment, 15 videos related to three genres, "movie", "news" and "sports", were prepared, where five videos were contained in each genre. The length of each video was 60 seconds. Five subjects watched these videos in the standing position at a place about two-meters away from a 15-inch display. The Kinect was set on the display to obtain the subjects' viewing behavior. The video features and their corresponding users' viewing behavior features were calculated every 0.5 seconds, and we obtained 1800 samples for the evaluation. Since it was difficult to obtain rating scores in such short periods, the subjects performed the ratings every ten seconds in five grades, i.e., five ordinal classes ($R=5$), after watching all of the videos. In this experiment, we adopted the above conditions for the simplicity of the experiment procedures.

For these datasets, we performed feature selection and predicted rating scores of videos based on SVOR (Support Vector Ordinal Regression) [10]. We conducted 15-fold cross-validation to compare the performance of our method with those of some comparative feature selection methods by using two evaluation metrics, MAE (Mean Absolute Error) and MZE (Mean Zero-one Error). It should be noted that we used the Gaussian kernel for SVOR, and its kernel parameter was determined based on the grid search [11] using MAE. Results of the final rating prediction are shown in Table 2. For comparisons, we adopted the benchmarking and state-of-the-art methods [3]–[5], where mRMR algorithm and mRMR-CCA algorithm are the feature selection methods only focusing on two modalities, (video,rating) or (video,behavior). In the mRMR algorithm [3], the mutual information was used.

The effectiveness of our method can be confirmed from the experimental results shown in Table 2. The direct comparison between the canonical correlation and the mutual information corresponds to the comparison between mRMR-CCA (video,rating) and mRMR in Table 2. Then the effectiveness of the use of the canonical correlation for the fea-

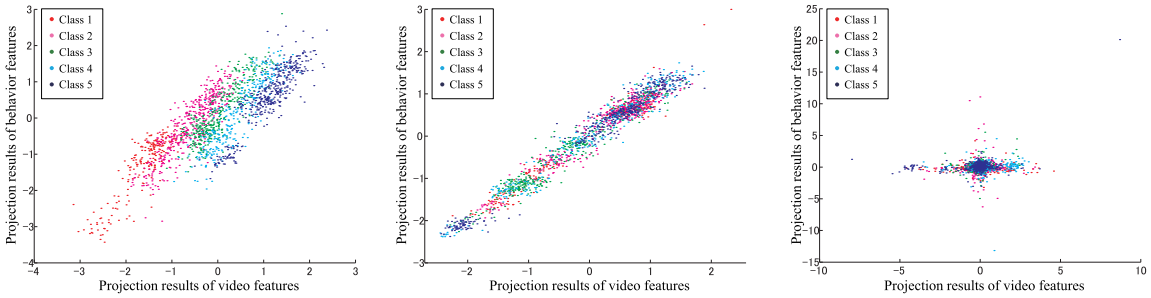

Fig. 1 Projection results of video features and viewing behavior features based on our method (mRMR-SCMMCCA algorithm), mRMR-CCA algorithm [4] and SLPCCA-OC algorithm [5]. The horizontal and vertical axes correspond to the projection results of video features and viewing behavior features, respectively.

ture selection can be verified. Furthermore, Fig. 1 shows projection results of video features and viewing behavior features obtained by our method (mRMR-SCMMCCA algorithm), the mRMR-CCA algorithm (video,behavior) and the SLPCCA-OC algorithm, where the corresponding graphs of the mRMR-CCA algorithm (video,rating) and the mRMR algorithm cannot be obtained since they only focus on the relationship between video features and rating scores. From Fig. 1, it can be seen that the proposed method provides the best projections that can separate the samples belonging to different classes, i.e., having different rating scores, based on the SCMMCCA.

4. Conclusions

A novel video feature selection algorithm (mRMR-SCMMCCA algorithm) for preference extraction has been presented in this paper. The experimental results have shown the superiority of our method and indicated that it becomes feasible to extract the users' individual preference as the selection of the optimal video features.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Numbers JP25280036 and JP15K12023.

References

[1] M. Haseyama, T. Ogawa, and N. Yagi, "A review of video retrieval based on image and video semantic understanding," ITE Trans. Media Technology and Applications, vol.1, no.1, pp.2–9, 2013.

[2] M. Takahashi, S. Clippingdale, M. Okuda, Y. Yamanouchi, M. Naemura, and M. Shibata, "An estimator for rating video contents on the basis of a viewer's behavior in typical home environments," 2013 International Conference on Signal-Image Technology Internet-Based Systems (SITIS), pp.6–13, 2013.

[3] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.8, pp.1226–1238, Aug. 2005.

[4] H. Kaya, F. Eyben, A.A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3729–3733, 2014.

[5] Y. Yamaguchi, T. Ogawa, S. Asamizu, and M. Haseyama, "Preference estimation for video recommendation from viewing behavior based on SLPCCA-OC," 2015 Joint Conference of the International Workshop on Advanced Image Technology (IWAIT) and the International Forum on Medical Imaging in Asia (IFMIA), 2015.

[6] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," International Conference on Digital Audio Effects, pp.237–244, 2007.

[7] H. Bay, T. Tuytelaars, and L. Gool, "SURF: Speeded up robust features," The 9th European Conference on Computer Vision, Lecture Notes in Computer Science, vol.3951, pp.404–417, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[8] J.F. Cohn, Z. Ambadar, and P. Ekman, Observer-based measurement of facial expression with the facial action coding system, pp.203–221, Oxford University Press, 2007.

[9] A.A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," IEEE Trans. Image Process., vol.11, no.3, pp.293–305, 2002.

[10] W. Chu and S.S. Keerthi, "New approaches to support vector ordinal regression," The 22nd International Conference on Machine Learning (ICML), pp.145–152, 2005.

[11] C.W. Hsu, C.C. Chang, and C.J. Lin, "A practical guide to support vector classification," Tech. Rep., Department of Computer Science, National Taiwan University, 2003.