PERSONALIZED VIDEO PREFERENCE ESTIMATION BASED ON EARLY FUSION USING MULTIPLE USERS' VIEWING BEHAVIOR

Yoshiki Ito[†]*, Takahiro Ogawa*[‡] *and Miki Haseyama*[‡]

[†]School of Engineering, Hokkaido University
N-13, W-8, Kita-ku, Sapporo, Hokkaido, 060-8628, Japan
[‡]Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: {ito, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper presents a novel method for personalized video preference estimation based on early fusion using multiple users' viewing behavior. The proposed method adopts supervised Multi-View Canonical Correlation Analysis (sMVCCA) to estimate correlation between different types of features. Specifically, we estimate optimal projections maximizing the correlation between three features of video, target user's viewing behavior and evaluation scores for video. Then novel video features (canonical video features), which reflect the target user's individual preference, are obtained by the estimated projections. Furthermore, our method computes sMVCCA-based canonical video features by using multiple users' viewing behavior and a target user's evaluation scores. This non-conventional approach using the multiple users' viewing behavior for the preference estimation of the target user is the biggest contribution of our method, and it enables early fusion of the canonical video features. Consequently, successful video recommendation that reflects the users' individual preference can be expected via the evaluation score prediction from the integrated canonical video features. Experimental results show the effectiveness of our method.

Index Terms— canonical correlation analysis, preference estimation, viewing behavior, early fusion.

1. INTRODUCTION

It has become easier to access a large amount of video via video-sharing services such as YouTube[1] or video-streaming services such as Hulu[2]. Generally, these services require users to input queries when the users retrieve their desired video. Thus, if users cannot provide suitable queries accurately reflecting their desired contents, successful retrieval of these contents becomes difficult [1]. To overcome this problem, many video recommendation methods that do not require such queries have been proposed. Most video recommendation methods are based on collaborative filtering and content-based filtering [2–5]. In the methods based on the collaborative filtering, the similar users who have similar preference are found for the target users on the basis of the evaluation scores of contents provided by the users. Although methods based on the collaborative filtering recommend video to target users based on similar users' evaluation scores, these methods cannot provide video that have not been evaluated in advance. Meanwhile, methods based on the content-based filtering recommend video by directly using their video features to solve the problem of the collaborative filtering. However, since they only monitor raw video features, it is difficult to effectively reflect target users' individual preference for the recommendation.

In general, even if different users provide the same evaluation scores for the same video, their individual preference for video may be different since each video contains several objects. Therefore, it is necessary to extract each user's individual preference, i.e., video features suitable for the target user. The study of feature selection has been carried out intensively, and many benchmarking algorithms have been proposed [6–11]. However, they can only monitor the relationship between video features and their corresponding evaluation scores which reflect preference degrees of video. Therefore, if different users provide the same evaluation scores for the same video, video features selected by these feature selection algorithms become the same, perfectly. Therefore, for extracting true preference of the target user, we need to use other elements, which are closely related to his/her preference.

In order to extract such preference, many methods use biological signals such as brain waves [12, 13], but this approach puts a burden on the users since most biological signals are obtained by an equipment that is attached to users' bodies [14, 15]. On the other hand, since in-cameras are mounted on a number of devices such as personal computers or smart-phones, acquisition of users' viewing behavior does not put a burden on the user. Since users' viewing behavior such as gazing, facial expression and body movements is closely related to the users' attention, it is one of the most important factors to extract the users' individual preference [16, 17]. Thus, there have been proposed several methods for predicting evaluation scores of video from the users' viewing behavior [17, 18]. Note that although video features that are closely related to each user's preference are different from each other, the existing methods do not consider this point.

In this paper, we present a novel method of video preference estimation for video recommendation. The proposed method enables derivation of new video features (canonical video features), which can reflect the individual preference, by estimation of projections maximizing the correlation between video, the target user's viewing behavior and evaluation scores. Supervised Multi-View Canonical Correlation Analysis (sMVCCA) [19] is utilized for estimating the projections, and the canonical video features that have the maximum correlation with the other two features can be obtained. Furthermore, the proposed method estimate sMVCCA-based canonical video features by using multiple users' viewing behavior and the target user's evaluation scores as shown in Fig 1. The use of the multiple users'

This work was partly supported by Grant-in-Aid for Scientific Research (B) JP25280036, Japan Society for the Promotion of Science (JSPS).

[1]http://www.youtube.com/

[2]http://www.hulu.com/

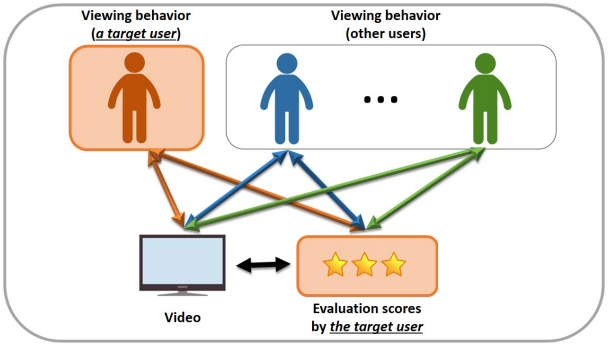

Fig. 1. The overview of our novel approach. Our method focuses on not only target user's viewing behavior but also other users' viewing behavior for extracting the target user's preference.

canonical correlations for the preference estimation of the target user is the biggest contribution of our method, and this enables early fusion of the canonical video features. Finally, Support Vector Ordinal Regression with Implicit Constraints (SVOR-IMC) [20] is trained with the canonical video features, and prediction of evaluation scores for new video becomes feasible. Consequently, realization of the video recommendation that can reflect the user's individual preference is expected. Note that our work shown in this paper is an extended version of [21, 22]. In this paper, we realize the new concept of the preference estimation from multiple users' viewing behavior.

2. EXTRACTION OF CANONICAL VIDEO FEATURES BASED ON EARLY FUSION

This section shows the extraction method of the canonical video features. First, in 2.1, we explain extraction of three kinds of features used in our method. Furthermore, the calculation of sMVCCA-based canonical video features based on the early fusion is presented in 2.2.

2.1. Extraction of Three Kinds of Features

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

Video features (1209 dimensions) :
As shown in Table 1, we adopt 145 audio features obtained by MIRToolbox [23], which consists of Dynamics, Spectral, Timbre, Tonal and Rhythm. Furthermore, HSV color histogram and Bag of visual words [24] based on Speeded-Up Robust Features (SURF) [25] are calculated as the visual features. Then the video feature vector $\boldsymbol{v}_i \in \mathbb{R}^{D_v}$ is obtained for each ith sample, where $D_v = 1209$ from Table 1.

Viewing behavior features (22 dimensions) :
By using a Kinect sensor[3], facial features and body movement features are obtained as shown in Table 1 . To calculate the facial features, we detect landmark points on the face, and a 3D face model corresponding to the landmark points is extracted by the Kinect. This 3D face model outputs head poses and facial expression descriptor based on Action Units [26], and the Microsoft Face Tracking Software Development Kit for Kinect for Windows (Face Tracking SDK) [4] supports six Action Units (Upper lip raiser, Jaw lowerer, Lip stretcher, Brow lowerer, Lip corner depressor, and Outer brow raiser). In this way, we can obtain the 14-dimensional facial features. Next, for calculating the body movement features, we obtain a user's region and coordinates of the user's skeleton from the Kinect. Then 8-dimensional body movement features are calculated as shown in Table 1. We then obtain the viewing behavior feature vector $\boldsymbol{b}_i \in \mathbb{R}^{D_b}$ for each ith sample, where $D_b = 22$ from Table 1.

Label features :
The target user evaluates video in R grades while watching them. We then obtain an evaluation score of ith sample of video, $l_i \in \{1, 2, \cdots, R\}$, from the target user. This score is converted into an R-dimensional label feature vector in 2.2.

The features extracted in the above procedures are obtained in the short time period corresponding to the frame rate of video. Given a constant time width s, we calculate the features of the time length $2s + 1$ for each ith sample. Specifically, we define the features $\boldsymbol{v}_i^s \in \mathbb{R}^{D_v}$, $\boldsymbol{b}_i^s \in \mathbb{R}^{D_b}$ and $l_i^s \in \mathbb{R}^{D_l}$ corresponding to the above three features as $\boldsymbol{v}_i^s = (\sum_{j=i-s}^{i+s} \boldsymbol{v}_j)/(2s+1)$, $\boldsymbol{b}_i^s = (\sum_{j=i-s}^{i+s} \boldsymbol{b}_j)/(2s+1)$ and $l_i^s = round\{(\sum_{j=i-s}^{i+s} l_j)/(2s+1)\}$, where $round\{\cdot\}$ indicates the rounding off an operator. These are trivial procedures, but contribute to the robust relationship estimation in the following subsection.

2.2. Extraction of sMVCCA-based Canonical Video Features using Early Fusion

This section presents the calculation method of the canonical video features that are derived on the basis of the relationship between video features, viewing behavior features and label features via sMVCCA and early fusion. First, we define a video feature matrix $\boldsymbol{V}^s \in \mathbb{R}^{D_v \times N}$ as $\boldsymbol{V}^s = [\boldsymbol{v}_1^s, \boldsymbol{v}_2^s, \cdots, \boldsymbol{v}_N^s]$. Next, we consider not only the target user's viewing behavior but also other users' viewing behavior in our method. Accordingly, we define a pth users' viewing behavior feature matrix $\boldsymbol{B}_p^s \in \mathbb{R}^{D_b \times N}$ as $\boldsymbol{B}_p^s = [\boldsymbol{b}_{1,p}^s, \boldsymbol{b}_{2,p}^s, \cdots, \boldsymbol{b}_{N,p}^s]$, where $p \in \{1, 2, \cdots, P\}$, and P is the number of users integrated by the early fusion, where we assume $p = 1$ corresponds to the target user. Furthermore, the target user's label feature matrix $\boldsymbol{L}^s \in \mathbb{R}^{D_l \times N}$ representing degrees of video preference are defined as $\boldsymbol{L}^s = [\boldsymbol{l}_1^s, \boldsymbol{l}_2^s, \cdots, \boldsymbol{l}_N^s]$. Note that $\boldsymbol{l}_i^s \in \mathbb{R}^{D_l}$ is a binary vector obtained from l_i^s based on [19], where $D_l = R$ as shown in Table 1. From these matrices, we try to solve the following optimization problem, which maximizes the sum of the three kinds of correlations, the correlation between $\boldsymbol{V}^s$ and $\boldsymbol{B}_p^s$ (video, viewing behavior), the correlation between $\boldsymbol{V}^s$ and $\boldsymbol{L}^s$ (video, label) and the correlation between $\boldsymbol{B}_p^s$ and $\boldsymbol{L}^s$ (viewing behavior, label) to obtain the optimal projections $\hat{\boldsymbol{w}}_{v,p}^s \in \mathbb{R}^{D_v}$, $\hat{\boldsymbol{w}}_{b,p}^s \in \mathbb{R}^{D_b}$ and $\hat{\boldsymbol{w}}_{l,p}^s \in \mathbb{R}^{D_l}$:

$$\{\hat{\boldsymbol{w}}_{v,p}^s, \hat{\boldsymbol{w}}_{b,p}^s, \hat{\boldsymbol{w}}_{l,p}^s\} = \arg \max_{\boldsymbol{w}_{v,p}^s, \boldsymbol{w}_{b,p}^s, \boldsymbol{w}_{l,p}^s} \{(\boldsymbol{w}_{v,p}^s)^T \boldsymbol{V}^s (\boldsymbol{B}_p^s)^T \boldsymbol{w}_{b,p}^s + (\boldsymbol{w}_{v,p}^s)^T \boldsymbol{V}^s (\boldsymbol{L}^s)^T \boldsymbol{w}_{l,p}^s + (\boldsymbol{w}_{b,p}^s)^T \boldsymbol{B}_p^s (\boldsymbol{L}^s)^T \boldsymbol{w}_{l,p}^s\}$$

$$\text{s.t.} \quad (\boldsymbol{w}_{v,p}^s)^T \boldsymbol{V}^s (\boldsymbol{V}^s)^T \boldsymbol{w}_{v,p}^s + (\boldsymbol{w}_{b,p}^s)^T \boldsymbol{B}_p^s (\boldsymbol{B}_p^s)^T \boldsymbol{w}_{b,p}^s + (\boldsymbol{w}_{l,p}^s)^T \boldsymbol{L}^s (\boldsymbol{L}^s)^T \boldsymbol{w}_{l,p}^s = 1. \tag{1}$$

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

[3] http://www.microsoft.com/en-us/kinectforwindows/

[4] http://msdn.microsoft.com/en-us/library/jj130970.aspx

Table 1. Three kinds of features used in the proposed method.

	Feature quantities	Dimensions
Video features	Sound (Dynamics, Spectral, Timbre, Tonal, Rhythm) [23]	145
	HSV color histogram	64
	Bag of visual words [24] based on SURF [25]	1000
Total	-	**1209**
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 Units [26]	6
Viewing behavior features (body)	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
	Angle of the body based on distance between both shoulders and a display	3
Total	-	**22**
Label features	Features based on the score evaluated by the user	R
Total	-	***R***

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{V}^s(\boldsymbol{B}_p^s)^T & \boldsymbol{V}^s(\boldsymbol{L}^s)^T \\ \boldsymbol{B}_p^s(\boldsymbol{V}^s)^T & \mathbf{0} & \boldsymbol{B}_p^s(\boldsymbol{L}^s)^T \\ \boldsymbol{L}^s(\boldsymbol{V}^s)^T & \boldsymbol{L}^s(\boldsymbol{B}_p^s)^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{v,p}^s \\ \boldsymbol{w}_{b,p}^s \\ \boldsymbol{w}_{l,p}^s \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{V}^s(\boldsymbol{V}^s)^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}_p^s(\boldsymbol{B}_p^s)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{L}^s(\boldsymbol{L}^s)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{v,p}^s \\ \boldsymbol{w}_{b,p}^s \\ \boldsymbol{w}_{l,p}^s \end{bmatrix}. \tag{2}$$

We then obtain the following projection matrix from the solution of the generalized eigenvalue problem:

$$\hat{\boldsymbol{W}}_{v,p}^s = [\hat{\boldsymbol{w}}_{v,p,1}^s, \hat{\boldsymbol{w}}_{v,p,2}^s, \cdots, \hat{\boldsymbol{w}}_{v,p,D_p}^s]. \tag{3}$$

Given that λ_d ($\lambda_d > \lambda_{d+1}; d = 1, 2, \cdots, D_p - 1$) are eigenvalues corresponding to λ in Eq. (2), $\hat{\boldsymbol{w}}_{v,p,d}^s$ are projection vectors corresponding to these eigenvalues, where $D_p < \min\{D_v, D_b, D_l\}$. We then integrate the projection matrices obtained from the P users' viewing behavior as follows:

$$\hat{\boldsymbol{W}}_v^s = [\hat{\boldsymbol{W}}_{v,1}^s, \hat{\boldsymbol{W}}_{v,2}^s, \cdots, \hat{\boldsymbol{W}}_{v,P}^s] \in \mathbb{R}^{D_v \times \hat{D}}, \tag{4}$$

where $\hat{D} = D_1 + D_2 + \cdots + D_P$. Furthermore, we calculate the canonical video features $\hat{\boldsymbol{V}} = [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_N] \in \mathbb{R}^{\hat{D} \times N}$, which reflect the target user's individual preference based on the early fusion as follows:

$$\hat{\boldsymbol{V}} = (\hat{\boldsymbol{W}}_v^s)^T \boldsymbol{V}^s. \tag{5}$$

In this way, we can calculate the new canonical video features from the multiple users' viewing behavior. The new canonical video features using the pth user's projection matrix $\hat{\boldsymbol{W}}_{v,p}^s$ support the prediction of the target user's evaluation scores for video as shown in the following section. In this way, we can obtain the non-conventional video features via the early fusion, which is the biggest contribution of our method.

3. SCORE PREDICTION BASED ON SVOR-IMC

In this section, we describe the method of evaluation score prediction for video recommendation by using SVOR-IMC [20]. In our method, SVOR-IMC is trained with the canonical video features obtained in the previous section. Given pairs of the canonical video feature vectors and the evaluation scores of training video $(\hat{\mathbf{v}}_i, l_i^s)(i = 1, 2, \cdots, N)$, the feature vectors $\hat{\mathbf{v}}_i$ are mapped into a high-dimensional Reproducing Kernel Hilbert Space (RKHS) to obtain $\phi(\hat{\mathbf{v}}_i)$. Note that $\hat{\mathbf{v}}_i \in \mathbb{R}^{\hat{D}}$ is the $\hat{D}$-dimensional vector of the canonical video features. The values $l_i^s \in \{1, 2, \cdots, R\}$ are the evaluation scores as shown in the previous section. The discrimination of different ordinal classes, i.e., prediction of an evaluation score, for a new test vector $\hat{\mathbf{v}}^{new}$ becomes feasible by the following discriminant function:

$$\arg \min_{j \in \{1,2,\cdots,R\}} \left\{ j : f(\hat{\mathbf{v}}^{new}) < \tau_j \right\}, \quad f(\hat{\mathbf{v}}^{new}) = \langle \boldsymbol{u} \cdot \phi(\hat{\mathbf{v}}^{new}) \rangle, \tag{6}$$

where $\langle \cdot \rangle$ denotes the inner product in the RKHS, $\boldsymbol{u}$ is a mapping direction, and τ_j are thresholds of class labels.

In order to obtain the optimal mapping direction and thresholds, the optimization problem of SVOR-IMC is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{u},\boldsymbol{\tau},\boldsymbol{\xi},\boldsymbol{\xi}^*} \ & \frac{1}{2} \langle \boldsymbol{u} \cdot \boldsymbol{u} \rangle + C \sum_{j=1}^{R-1} \left(\sum_{k=1}^{j} \sum_{i=1}^{N^k} \xi_{ki}^j + \sum_{k=j+1}^{R} \sum_{i=1}^{N^k} \xi_{ki}^{*j} \right) \\ \text{s.t.} \quad & \left\langle \boldsymbol{u} \cdot \phi(\hat{\mathbf{v}}_i^k) \right\rangle - \tau_j \le -1 + \xi_{ki}^j, \ \xi_{ki}^j \ge 0, \\ & \text{for } k = 1, \cdots, j \ \text{ and } \ i = 1, \cdots, N^k; \\ & \left\langle \boldsymbol{u} \cdot \phi(\hat{\mathbf{v}}_i^k) \right\rangle - \tau_j \ge 1 - \xi_{ki}^{*j}, \ \xi_{ki}^{*j} \ge 0, \\ & \text{for } k = j+1, \cdots, R \ \text{ and } \ i = 1, \cdots, N^k, \end{aligned} \tag{7}$$

where $C > 0$ is a constant variable, N^k denotes the number of samples in class k, $\hat{\mathbf{v}}_i^k \in \mathbb{R}^{\hat{D}}$ is the canonical video feature vector belonging to class k, and ξ_{ki}^j and ξ_{ki}^{*j} are slack variables. The details of the slack variables ξ_{ki}^j and ξ_{ki}^{*j} are shown in [20]. Then the dual problem of Eq. (7) is derived by using the Lagrange multiplier approach:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \ & -\frac{1}{2} \sum_{k,i} \sum_{k',i'} \left(\sum_{j=1}^{k-1} \alpha_{ki}^{*j} - \sum_{j=k}^{R-1} \alpha_{ki}^j \right) \left(\sum_{j=1}^{k'-1} \alpha_{k'i'}^{*j} - \sum_{j=k'}^{R-1} \alpha_{k'i'}^j \right) \mathcal{K}(\hat{\mathbf{v}}_i^k, \hat{\mathbf{v}}_i^{k'}) \\ & + \sum_{k,i} \left(\sum_{j=1}^{k-1} \alpha_{ki}^{*j} + \sum_{j=k}^{R-1} \alpha_{ki}^j \right) \\ \text{s.t.} \quad & \sum_{k=1}^{j} \sum_{k=i}^{N^k} \alpha_{ki}^j = \sum_{k=j+1}^{R} \sum_{i=1}^{N^k} \alpha_{ki}^{*j} \ \ \forall j, \\ & 0 \le \alpha_{ki}^j \le C \quad \forall j \text{ and } k \le j, \\ & 0 \le \alpha_{ki}^{*j} \le C \quad \forall j \text{ and } k > j, \end{aligned} \tag{8}$$

Table 2. Quantitative evaluation of the proposed method and the comparative methods.

	Proposed Method		Method 1		Method 2		Method 3		Method 4		Method 5	
Subject	MAE	MZE	MAE	MZE	MAE	MZE	MAE	MZE	MAE	MZE	MAE	MZE
1	**0.708**	0.523	0.738	**0.498**	0.755	0.542	0.754	0.531	1.062	0.763	1.098	0.849
2	**0.562**	**0.473**	0.619	0.496	0.647	0.519	0.688	0.524	0.849	0.696	0.900	0.766
3	**0.708**	**0.532**	0.771	0.551	0.736	0.534	0.844	0.601	1.028	0.641	1.102	0.738
4	**1.101**	**0.594**	1.172	0.605	1.181	0.660	1.228	0.664	1.523	0.777	1.892	0.837
5	**0.510**	**0.419**	0.542	0.428	0.528	0.436	0.551	0.487	0.706	0.579	0.710	0.584
Average	**0.718**	**0.508**	0.768	0.516	0.769	0.538	0.813	0.561	1.034	0.691	1.140	0.755

where α_{ki}^{j} and α_{ki}^{*j} are the Lagrangian multipliers, and $\mathcal{K}(\hat{\mathbf{v}}, \hat{\mathbf{v}}') = \langle \phi(\hat{\mathbf{v}}) \cdot \phi(\hat{\mathbf{v}}') \rangle$. By using the optimal α_{ki}^{j} and α_{ki}^{*j}, the discriminant function for a new input vector $\hat{\mathbf{v}}^{new}$ in Eq. (6) can be rewritten as

$$\begin{aligned} \arg \min_{j \in \{1,2,\cdots,R\}} & \quad \left\{ j : f(\hat{\mathbf{v}}^{new}) < \tau_j \right\} \\ f(\hat{\mathbf{v}}^{new}) & = \sum_{k,i} \left(\sum_{j=1}^{k-1} \alpha_{ki}^{*j} - \sum_{j=k}^{R-1} \alpha_{ki}^{j} \right) \mathcal{K}(\hat{\mathbf{v}}_i^k, \hat{\mathbf{v}}^{new}). \end{aligned} \quad (9)$$

Consequently, we can perform the prediction of the target user's evaluation scores via SVOR-IMC from the canonical video features.

4. EXPERIMENTAL RESULTS

In this section, we show experimental results to verify the effectiveness of our method. First, in this experiment, three keywords "movie", "news" and "sports" were given as queries to YouTube. Then five video per each keyword were obtained, and these 15 video were used for the experiment. The subjects were five healthy men who were approximately 22 years old, and they watched 15 video for 65 seconds each in the sitting position at a place about 1.5 meters away from the display. Note that we do not extract features for five seconds immediately after watching each video to avoid noise by the user. The Kinect was set on a 15-inch display to obtain the subjects' viewing behavior. The subjects then evaluated all video by a value of five ordinal grades, i.e., 5 (very favorite), 4 (favorite), 3 (undecided), 2 (unfavorite) and 1 (unfavorite at all), by console input using the keyboard in real time. Therefore, the dataset including the three features (video, viewing behavior, label) can be obtained.

Next, we explain the experimental conditions. In this experiment, we adopted the Gaussian kernel $\mathcal{K}(\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j) = \exp\{-\|\hat{\mathbf{v}}_i - \hat{\mathbf{v}}_j\|^2 / 2\sigma^2\}$ for SVOR-IMC, where the kernel width σ^2 was chosen by searching the following parameter space: $\sigma^2 \in [2^{-15}, 2^{-13}, 2^{-11}, \cdots, 2^3]$. Additionally, the constant variable C in Eq. (7) was chosen by searching the following parameter space: $C \in [2^{-5}, 2^{-3}, 2^{-1}, \cdots, 2^7]$. We then decided these parameters by grid search [27]. Moreover, we determined the dimension of the projection matrices from $D_p \in [1, 2, 3, 4]$. In the proposed method, the number of users to integrate projections in Eq. (4) was set as $P = 5$ since this number corresponds to the number of subjects. In addition, time width s in 2.1 was simply set to one second.

In this experiment, we conducted 15-fold cross-validation to compare the performance of our method with those of some comparative methods by using the following score metrics:

$$\text{MAE} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| l_i^{PRE} - l_i^{GT} \right|, \quad \text{MZE} = \frac{1}{N_t} \sum_{i=1}^{N_t} 1_{l_i^{PRE} \neq l_i^{GT}},$$

where N_t is the number of test samples, l_i^{PRE} is a predicted evaluation score of ith sample, and l_i^{GT} is the ground-truth score of ith sample. Furthermore, $1_{l_i^{PRE} \neq l_i^{GT}}$ outputs one when $l_i^{PRE} \neq l_i^{GT}$ is satisfied. Otherwise, it outputs zero. Specifically, the lower both of these scores are, the higher the accuracy of the method is. In this experiment, we determined each parameter to the value corresponding to the best MAE.

Results of our experiment are shown in Table 2. In this table, we also show the results of five comparative methods. We show the method considering only the time width shown in 2.1 not using the multiple user integration as Method 1, the method using only the multiple user integration not considering the time width as Method 2. Moreover, we show our previously reported method [22] as Method 3, which is one of the state-of-the-art methods. In addition, we show the method using SLPCCA [28] as Method 4. Finally, we show the method using the standard CCA [29] as Method 5. By comparing the results between "Methods 3 and 4" and "Method 5", we can confirm the effectiveness of using viewing behavior. Moreover, by comparing results between the proposed method and the other methods except Methods 4 and 5, we can confirm the effectiveness of the proposed method which utilizes not only the target user's viewing behavior but also the multiple users' viewing behavior.

As shown in Table 2, the results show the effectiveness of the propopsed method since we can see the MAE and MZE of our method are almost lower than those of the other comparative methods. First, by comparing Methods 3 and 4 with Method 5, it is confirmed that utilizing user's viewing behavior is highly effective. Next, by comparing Methods 1 and 2 with Method 3, it is confirmed that considering the time width for each feature and the integration based on the multiple users' viewing behavior are effective. Furthermore, it is effective to consider both approaches by comparing the proposed method with Methods 1 and 2. We can also see that the projections utilizing the multiple users' viewing behavior can support preference estimation of the target user compared to those utilizing only the target user's viewing behavior. Thus, the proposed method can extract the video features that reflect the users' individual preference more successfully.

5. CONCLUSIONS

The method for video preference estimation for video recommendation has been presented in this paper. The proposed method newly introduces multiple users' viewing behavior to calculate the projections which provide the target user's preference more accurately. This is the biggest contribution of this paper, and the experimental results have shown the superiority of our method. The realization of video recommendation that reflects the users' individual preference can be achieved.

6. REFERENCES

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

[2] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5-53, 2004.

[3] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005.

[4] H. S. Tan and H. W. Ye, "A collaborative filtering recommendation algorithm based on item classification," in Proceedings of *Pacific-Asia Conference on Circuits, Communications and System (PACCS)*, 2009, pp. 694-697.

[5] H. Li, F. Cai, and Z. Liao, "Content-based filtering recommendation algorithm using HMM," in Proceedings of *International Conference on Computational and Information Sciences (ICCIS)*, 2012, pp. 275-277.

[6] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.

[7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[8] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189-201, 2009.

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

[10] V. Bolong-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483-519, 2013.

[11] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[12] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3498-3510, 2012.

[13] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "EEG-based classification of music appraisal responses using time frequency analysis and familiarity ratings," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 161-172, 2013.

[14] A. S. AlMejrad, "Human emotions detection using brain wave signals: A challenging," *Europian Journal of Scientific Research*, vol. 44, no. 4, pp. 640-659, 2010.

[15] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in Proceedings of *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 827-834.

[16] M. I. Posner, "Orienting of attention," *The Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3-25, 1980.

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

[18] T. Shiraishi, T. Ogawa, and M. Haseyama, "Video recommendation using user interest extracted from viewing behavior," in Proceedings of *The 27th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2012, pp. PM2-13.

[19] G. Lee, A. Singanamalli, H. Wang, M. D. Feldman, S. R. Master, N. N. C.Shih, E. Spangler, T. Rebbeck, J. E. Tomaszewski, and A. Madabhushi, "Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 284-297, 2015.

[20] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Computation*, vol. 19, no. 3, pp. 792-815, 2007.

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

[22] Y. Ito, T. Ogawa, and M. Haseyama, "Novel video feature-based favorite video estimation using users' viewing behavior and evaluation," in Proceedings of *2016 IEEE Global Conference on Consumer Electronics (GCCE)*, 2016, pp. 224-225.

[23] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in Proceedings of *International Conference on Digital Audio Effects (DAFX)*, 2007, pp. 237-244.

[24] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in Proceedings of *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, vol. 1, pp. 1-2.

[25] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008.

[26] J. F. Cohn, Z. Ambadar, and P. Ekman, "Observer-based measurement of facial expression with the facial action coding system," *The Handbook of Emotion Elicitation and Assessment*, 2007, pp. 203-221.

[27] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," *Technical Report, Department of Computer Science, National Taiwan University*, 2003.

[28] J. Yang and X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 623-628, 2012.

[29] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321-377, 1936.