This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2702749, IEEE Transactions on Affective Computing

Content-based Video Emotion Tagging Augmented by Users' Multiple Physiological Responses

Shangfei Wang,Shiyu Chen,Qiang Ji

Abstract—The intrinsic interactions among a video's emotion tag, its content, and a user's spontaneous responses while consuming the video can be leveraged to improve video emotion tagging, but such interactions have not been thoroughly exploited yet. In this paper, we propose a novel content-based video emotion tagging approach augmented by users's multiple physiological responses, which are only required during training. Specifically, a better emotion tagging model is constructed by introducing similarity constraints on the classifiers from video content and multiple physiological signals available during training. Maximum margin classifiers are adopted and efficient learning algorithms of the proposed model are also developed. Furthermore, the proposed video emotion tagging approach is extended to utilize incomplete physiological signals, since these signals are often corrupted by artifacts. Experiments on four benchmark databases demonstrate the effectiveness of the proposed method for implicitly integrating multiple physiological responses, and its superior performance to existing methods using both complete and incomplete multiple physiological signals.

Index Terms—Video emotion tagging, EEG signals, peripheral physiological signals, support vector machine

1 Introduction

Video emotion tagging has attracted increasing attention in recent years due to its wide potential applications in video creation and distribution. Automatic video content analysis and annotation is necessary to organize videos effectively and assist users in finding videos quickly. Emotion is an important component in the classification and retrieval of digital videos. Current video emotion tagging can be divided into two approaches [1]: direct approaches and implicit approaches. Direct video emotion tagging assigns emotion tags to videos based on users' examination of the video content. Implicit video emotion tagging, on the other hand, infers videos' emotion tags from a user's spontaneous nonverbal responses while consuming the videos [2].

Emotion is subjective by nature and involves physiological changes **in response to a stimulus**; therefore, videos' emotion tags are intrinsically linked to the video content and users' spontaneous responses. Fully exploiting the relationships among video content, users' spontaneous responses, and emotional descriptors will reduce the semantic gap between the low-level audio-visual features and the users' high-level emotional descriptors. However, current direct approaches only map from video content to emotional descriptors, and implicit approaches only explore mapping from users' spontaneous responses to emotional descriptors. Little research considers the relationships among video content, users' spontaneous responses, and emotional descriptors simultaneously [3] [4]. Furthermore, implicit methods typically combine video content and users' responses by explicitly fusing them during both training and testing. Although the development of wearable devices like smartwatches/bracelets allows us to capture some physiological signals, such as skin temperature (TEMP), other physiological signals, such as electroencephalograph (EEG) and electro-oculogram (EOG) readings, can only be recorded by expensive and complex sensors. Furthermore, users may dislike wearing sensors to record their physiological changes during the actual tagging, it is more practical to employ only video content without any user interaction during actual video tagging. To resolve this conflict, we propose an implicit fusion approach which only requires user's physiological responses during training and employs video content during actual tagging [5] [6].

We propose a novel content-based video emotion tagging approach augmented by users' multiple physiological responses, which are only required during training. First, features are extracted from multiple physiological signals and video content. For physiological signals, different frequency-domain and time-domain features are extracted from different signals. Audio-visual features are extracted from the video content. Then, we use similarity constraints on the mapping functions to capture the relationships among users' multiple physiological features, audio-visual features, and emotional descriptors during training. After that, we obtain a better video emotion classifier from video content with the help of multiple physiological signals. During testing, only video content is required. Maximum margin classifiers are adopted, and efficient learning algorithms of the proposed model are also developed. Furthermore, we extend the proposed video emotion tagging approach to utilize incomplete physiological signals, since physiological signals may be unavailable for any number of reasons. Instead of discarding the whole data instance if only a part is corrupted, we keep all available videos and physi-

- *Shangfei Wang is the corresponding author. Shangfei Wang and Shiyu Chen are with the School of Computer Science and Technology, University of Science and Technology of China, 230027, Hefei, China.*
E-mail: sfwang@ustc.edu.cn; sy1001@mail.ustc.edu.cn
- *Qiang Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, NY 12180-3590.*
E-mail: qji@ecse.rpi.edu

ological signals to train the emotion classifiers, and employ all available physiological signals to improve content-based video emotion tagging during training. Thus, we avoid a substantial amount of unusable data without discarding the good portion of the data. Experiments on the DEAP database [7], the MAHNOB-HCI database [8], the USTC-ERVS database [9], and the LIRIS-ACCEDE database [10] demonstrate the superiority of our proposed method to the existing methods, not only for video emotion tagging but also in the implicit integration of multiple physiological responses.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work on video emotion tagging. In section 3, we present the framework and details of our method. Section 4 shows the experimental results on four benchmark databases, as well as the analyses and the comparison to current work. Section 5 concludes the paper.

2 Related work

The concept of computational media aesthetics (CMA) proposed by Chitra Dorai [11] may be the first work on emotion tagging of videos. The purpose of CMA is to interpret audiences' emotional responses to media from visual and aural elements based on the film grammar. Later, Hanjalic and Xu [12] successfully linked the arousal and valence dimensions to low-level audio-visual features extracted from videos. This kind of emotion tagging research uses the direct approach, which classifies emotions from the related audio-visual features.

In addition to visual and aural elements in videos, users' spontaneous nonverbal responses while consuming the videos provide clues to actual emotions induced by the videos, and therefore provide indirect characterization of the video's emotion content. Some researchers have assigned emotion tags to videos based on emotions auto-recognized from users' spontaneous nonverbal response. This is called the implicit approach [13]. A comprehensive survey of video emotion tagging can be found in [1].

Since emotion is subjective by its nature and involves physiological changes **in response to a stimulus, video emotion tagging should involve video content and users' spontaneous responses**. However, few researchers have fully explored the relationship between them [4]. Soleymani et al. [3] adopted a linear relevance vector machine to analyze the relationship between subjects' physiological responses and video's emotion tags, as well as the relationship between the video content and the emotion tag. Their experimental results demonstrate that there is a significant correlation between emotion tags and physiological responses as well as between emotion tags and video content. Wang et al. [4] were among the first to combine users' EEG signals and video content for emotion annotation. They constructed three Bayesian Networks (BNs) to annotate videos by combining the video and EEG features at independent feature-level fusion, decision-level fusion, and dependent feature-level fusion. Their experimental results prove that the fusion methods outperform conventional emotion tagging methods that use video or EEG features alone. Moreover, the semantic gap between the low-level audio-visual features and the users' high-level emotion tags can be narrowed down with the help of EEG features. A downside of their approach is that users' EEG signals are required during both training and testing. We refer this method as explicit hybrid video emotion tagging.

While physiological signals are important for video emotion tagging, it is inconvenient to collect users' physiological signals during actual emotion tagging due to the high cost of physiological sensors and the comfort of users. Wang et al. [5] proposed a video tagging method with the aid of EEG signals, which are only available during training, but not available during testing. Specifically, a new video feature space is constructed using canonical correlation analysis (CCA) with the help of video content. A support vector machine (SVM) is adopted as the classifier on the constructed video feature space. Han et al. [14] proposed to recognize arousal levels by integrating low-level audio-visual features derived from video content and the human brain's functional activity in response to videos as measured by functional magnetic resonance imaging (fMRI). Specifically, a joint representation is learned by integrating video content and fMRI data using a multi-modal deep Boltzmann machine (DBM). The learned joint representation is utilized as the feature for training classifiers. The DBM fusion model can predict the joint representation of the videos without fMRI scans. Both studies integrate video content and users' physiological response to learn a new representation of videos during training. The new representation is more discriminative for emotion tagging than using each modality alone. Physiological data are only required during training to construct better video representation. During testing, only videos are available. We refer to this as implicit hybrid video emotion tagging, which is a more practical approach than explicit hybrid video emotion tagging.

We prefer to incorporate physiological response implicitly, so that these responses are only needed during training. Although the constructed representations in [5] and [14] reflect the intrinsic relationship between video content and users' physiological responses, they have no direct relationship to target emotion tags. To address this, we propose a new implicit hybrid video emotion tagging approach, which utilizes physiological responses to directly construct a better classifier for emotion tagging. Instead of using one kind of physiological signal like the two studies in [5] and [14], our approach can integrate multiple physiological signals to facilitate video emotion tagging. We impose the constraints on the mapping functions from video content, EEG features, and features of peripheral physiological signals to model the relationships among them. We modify SVM with our assumptions to train a better classifier from audio-visual features during training for better performance with the help of the other two kinds of features.

Current hybrid video emotion tagging research assumes that all channels of data, including videos and users' physiological signals, are always available. However, missing or corrupt data is common when investigating physiological signals. Physiological signals may be corrupted by power line interference, motion artifacts, electrode contact noise, or sensor device failure. To the best of our knowledge, there is little work on emotion tagging that considers missing or corrupted physiological data. This situation is the same in the field of emotion recognition from multiple physiological

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2702749, IEEE Transactions on Affective Computing

signals. Researchers have only recently begun to realize that it is too optimistic to assume that all data from all modalities is available at all times. Wagner et al. [15] may be the first to explore fusion methods for multi-modal emotion recognition with missing data. Other than discarding all data instances containing invalid modalities, which results in a substantial amount of unusable data, they propose to handle missing data at the decision-level fusion by integrating all the available modalities. Similarly, in this paper we extend the proposed content-based video emotion tagging approach to handle missing physiological data by integrating all the available modalities. Specifically, we employ all the available data during training to jointly train the emotion classifier from videos and the emotion classifiers from physiological signals. We use similarity constraints on these emotion classifiers to capture the intrinsic relationships among emotion labels, video content, and users' multiple physiological responses, to improve video emotion tagging.

This paper is an extension of our previous work [6]. In [6], we propose an implicit hybrid video emotion tagging approach that integrates video content and users' multiple physiological responses, which are only required during training. Specifically, we modify SVM with similarity constraints on classifiers to improve video emotion tagging, and conduct tagging experiments on three benchmark databases, i.e., the DEAP database, the MAHNOB-HCI database, and the USTC-ERVS database. Compared to our previous work, in this paper, we have extended our proposed content-based video emotion tagging method for incomplete physiological data, and conducted video emotion tagging experiments with complete physiological signals on the LIRIS-ACCEDE database and incomplete physiological signals on two databases, i.e., the DEAP database and the MAHNOB-HCI database.

3 Proposed Method

Our goal is to develop a method for video emotion tagging in which three different feature spaces can be obtained from training samples, and only one is required for testing samples. Specifically, we extract audio-visual features, EEG features, and features of peripheral physiological signals for training samples and only use audio-visual features for testing samples. **Peripheral physiological signals include electro-oculogram readings (EOG), electromyograms of zygomaticus and trapezius muscles (EMG), electrocardiograph (ECG), galvanic skin response (GSR), respiration amplitude (RSP), skin temperature (TEMP), and blood volume by plethysmograph signals (BVP)**. The framework of our proposed model is summarized in Fig.1. The details are described in the following subsections.

3.1 Feature Extraction

We extract audio-visual features, EEG features, and features of peripheral physiological signals for training samples and use only audio-visual features for testing samples.

3.1.1 EEG Features

First, noise reduction is carried out. A band-pass filter with a lower cutoff frequency of 0.3Hz and a higher cutoff frequency of 45Hz is adopted to remove the DC drifts and suppress the 50Hz power line interference [16] [17]. Then the spectral power from theta (4Hz $< f <$ 8Hz), slow alpha (8Hz $< f <$ 10Hz), alpha (8Hz $< f <$ 12Hz), beta (12Hz $< f <$ 30Hz), and gamma (30Hz $< f$) bands are extracted from all 32 electrodes as features. In addition to power spectral features, the difference between the spectral power of all the symmetrical pairs of electrodes on the right and left hemisphere is extracted to measure possible asymmetry in the brain activities due to emotional stimuli [7] [8]. The total number of EEG features for 32 electrodes is 216.

3.1.2 Features of Peripheral Physiological Signals

Peripheral physiological signals include EOG, EMG, ECG, GSR, RSP, TEMP, and BVP signals. Before feature extraction, these signals are preprocessed using band-pass filters to restrain the noise. Then, several commonly used features are adopted.

For EOG and EMG signals, 0.4Hz and 1Hz low-pass filters are adopted respectively. Energy, mean, and variance are extracted from 4 electrodes as features [7]. There are 12 EOG features and 12 EMG features.

For ECG signals, a 1Hz low-pass filter is used. Heart rate variability (HRV), root mean square of the mean squared difference of successive beats, standard deviation of beat interval change per respiratory cycle, 14 spectral power in the bands from 0-1.5Hz, low frequency 0.01-0.08Hz, medium frequency 0.08-0.15Hz and high frequency 0.15-0.5Hz components of HRV power spectrum, and Poincare analysis features (2 features) [8] [18] are extracted as features. The total number of ECG features is 22.

For GSR signals, mean, mean of the derivative, mean of the positive derivatives, proportion of negatives in the derivative, number of local minima, and 10 spectral powers within 0-2.4Hz [7] [8] are extracted as features after using a 3Hz low-pass filter. The total number of GSR features is 15.

For RSP signals, a 3Hz low-pass filter is adopted. Band energy ratio, average respiration signal, mean of the derivative, standard derivation, range of greatest breath, 10 spectral powers within 0-2.4Hz, average and median peak to peak time are extracted as features [7] [8]. The total number of RSP features is 17.

For TEMP signals, mean, mean of the derivative, spectral powers in 0-0.1 Hz and 0.1-0.2 Hz are extracted as features after a 0.45Hz low-pass filter is used [7] [18]. The total number of TEMP features is 4.

For BVP signals, 0.5Hz low-pass filter is adopted. BVP signals can be used to compute the HRV. Average and standard derivation of HRV and inter-beat intervals; energy ratio between 0.04-0.15 Hz and 0.15-0.5 Hz; and spectral power in 0.1-0.2 Hz, 0.2-0.3 Hz, 0.3-0.4 Hz, 0.01-0.08 Hz, 0.08-0.15 Hz, and 0.15-0.5 Hz components of HRV are used as features [7]. The total number of BVP features is 11.

3.1.3 Audio-Visual Features

Audio and visual features are extracted from video content. For audio features, 31 commonly used features including average energy, average loudness, spectrum flux, zero crossing rate (ZCR), standard deviation of ZCR, 12 Mel-frequency Cepstral Coefficients (MFCCs), log energy (as a kind of MFCC), and the standard deviations of the above 13 MFCCs

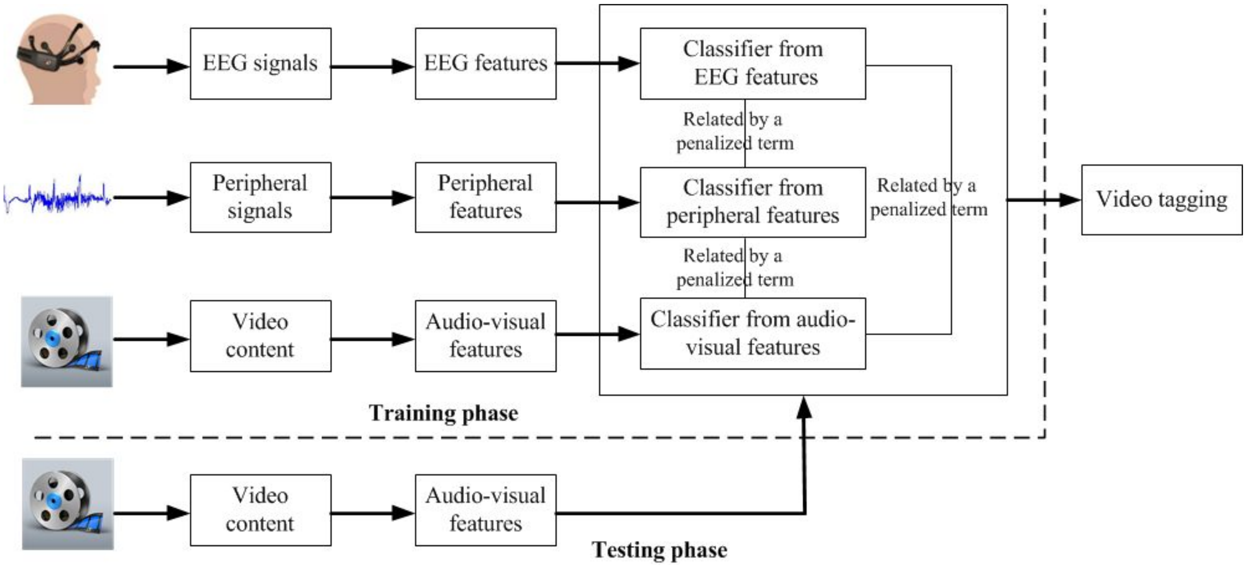
EEG signals
EEG features
Classifier from EEG features
Related by a penalized term
Peripheral signals
Peripheral features
Classifier from peripheral features
Related by a penalized term
Related by a penalized term
Video tagging
Video content
Audio-visual features
Classifier from audio-visual features
Training phase
Video content
Audio-visual features
Testing phase

Fig. 1. The framework of our approach

[19] are extracted using PRAAT(V5.2.35) [5]. For visual features, lighting key, color energy, and visual excitement are extracted from video clips [20].

3.2 Content-based Video Emotion Tagging Augmented by Multiple Physiological Signals

The proposed content-based video emotion tagging method is based on the SVM classifier. We enhance video emotion tagging by integrating the basic SVM with the similarity constraints among the mappings of video content and multiple physiological signals to emotion labels.

3.2.1 Problem Statement

In this paper we concentrate on the three-view case. All the features extracted from one modality construct one feature space. There are three different feature spaces in our training samples. Each one can be used to train a model. Instead of training feature spaces separately, we combine them with the similarity constraints. The objective of video emotion tagging is to train a classifier that maps the features of a sample into its true tag.

Audio-visual features, EEG features, and features of peripheral physiological signals are denoted as $v \in R^{|V|}$, $e \in R^{|E|}$, and $p \in R^{|P|}$ respectively, where $|V|$, $|E|$, and $|P|$ are the dimension of each feature space. Denote the emotion tag associated with a sample as $y \in \{-1, 1\}$. Training data consists of features in three feature spaces and emotion tags, denoted as $D = \{v_i, e_i, p_i, y_i | i = 1, ..., l\}$, where l is the number of training samples. During the testing phase, only **audio-visual** features are used, denoted as $T = \{v_i | i = 1, ..., t\}$, where t is the number of testing samples.

Our training samples D can be split into three parts:

$$DS = \{(v_i, y_i), (e_i, y_i), (p_i, y_i) | i = 1, ..., l\} \qquad (1)$$

which can be used to build three different mappings: f_v, f_e, and f_p respectively. These spaces provide different description powers, but they are all beneficial to video emotion tagging. Motivated by the work in [21], we use the similarity constraint between three distinct SVMs, each trained from one view of the data, to improve the performance of the related classifier. In this way, even if there is only one feature space available during testing, the information of the other feature spaces of training data can still be used on the tagging mission. The similarity constraint of $\{f_v, f_e\}$ can be represented as follows:

$$|f_v(v_i) - f_e(e_i)| \leq \eta_i^{ve} + \epsilon \qquad (2)$$

where η_i^{ve} is the slack variable to measure the amount that i^{th} sample fails to meet ϵ similarity between mapping f_v and f_e. This is an ϵ-intensity 1-normal constraint. The similarity constraints of $\{f_v, f_p\}$ and $\{f_e, f_p\}$ are identical with the similarity constraint of $\{f_v, f_e\}$. The similarity constraints of the mapping of $\{f_v, f_e\}$, $\{f_v, f_p\}$, and $\{f_e, f_p\}$ are combined into the three SVMs to enhance the effect of the classifier.

3.2.2 Support Vector Machine

The SVM is the basic classifier of our method. The SVM classifier first maps the feature into a real value. Given input feature x, the mapping is defined as:

$$f_x(x, \theta) = < w_x, \phi(x) > + b_x \qquad (3)$$

where ϕ is a kernel that maps input feature space into the kernel space. $\theta = \{w_x, b_x\}$ are parameters of the model.

Then SVM can be viewed as a 1-dimensional mapping followed by a decision function. The decision function is:

$$d(x) = sign(f_x(x)). \qquad (4)$$

Given training data $\{x_i, y_i\}$, SVM learns the model parameters by solving the following optimization problem:

$$
\begin{aligned}
&\underset{\Theta}{Min}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\
&\mathbf{s.t.} \\
&\qquad y_i(< w, \phi(x_i) > +b) \geq 1 - \xi_i, \\
&\qquad \xi_i \geq 0 \\
&\qquad all\ for\ i = 1, ..., l
\end{aligned}
\tag{5}
$$

where ξ_i is the slack variable which allows for the misclassification of some samples. The solution to Eq. (5) can be obtained by solving its dual problem.

3.2.3 Support Vector Machines with Similarity Constraints

Combining the three-view constraints with the SVMs, we obtain the following optimization problem of the proposed model:

$$
\begin{aligned}
&\underset{\Theta}{Min}\ \frac{1}{2}\|w_v\|^2 + \frac{1}{2}\|w_e\|^2 + \frac{1}{2}\|w_p\|^2 \\
&\qquad + C^v\sum_{i=1}^{l}\xi_i^v + C^e\sum_{i=1}^{l}\xi_i^e + C^p\sum_{i=1}^{l}\xi_i^p \\
&\qquad + D^{vp}\sum_{i=1}^{l}\eta_i^{vp} + D^{ve}\sum_{i=1}^{l}\eta_i^{ve} + D^{pe}\sum_{i=1}^{l}\eta_i^{pe} \\
&\mathbf{s.t.} \\
&\qquad |f_v(v_i) - f_p(p_i)| \leq \eta_i^{vp} + \epsilon, \\
&\qquad |f_v(v_i) - f_e(e_i)| \leq \eta_i^{ve} + \epsilon, \\
&\qquad |f_p(p_i) - f_e(e_i)| \leq \eta_i^{pe} + \epsilon, \\
&\qquad y_i f_v(v_i) \geq 1 - \xi_i^v, \\
&\qquad y_i f_e(e_i) \geq 1 - \xi_i^e, \\
&\qquad y_i f_p(p_i) \geq 1 - \xi_i^p, \\
&\qquad \xi_i^v \geq 0,\ \xi_i^e \geq 0,\ \xi_i^p \geq 0, \\
&\qquad \eta_i^{vp} \geq 0,\ \eta_i^{ve} \geq 0,\ \eta_i^{pe} \geq 0 \\
&\qquad all\ for\ i = 1,\ ...,\ l
\end{aligned}
\tag{6}
$$

where $\Theta = \{w_v, w_e, w_p, b_v, b_e, b_p\}$ are the parameters to be optimized. $\{C^v, C^e, C^p, D^{vp}, D^{ve}, D^{pe}\}$ are the weighted coefficients. The mapping functions f are Eq. (3) by using v, e, and p as the input features respectively. The first six terms of Eq. (6) are three object and slack terms which are the same as that in SVM: each pair for a feature space. The last three terms are the new terms, which are slack variables measuring any two of the three mappings that fail to meet ϵ similarity. With the new terms, the mappings f_v, f_e, and f_p are constrained to map the related feature spaces into similar 1-dimensional spaces.

This optimization problem can be solved by applying Lagrange multiplier techniques. We arrive at the following dual problem:

$$
\begin{aligned}
&\underset{\alpha}{Max}\ -\frac{1}{2}\sum_{i,j=1}^{l}[g_i^v g_j^v K_v(x_i, x_j) + g_i^e g_j^e K_e(x_i, x_j) + \\
&\qquad g_i^p g_j^p K_p(x_i, x_j)] + \sum_{i=1}^{l}(\alpha_i^v + \alpha_i^e + \alpha_i^p) \\
&\mathbf{s.t.} \\
&\qquad g_i^v = \alpha_i^v * y_i - \beta_i^{vp+} + \beta_i^{vp-} - \beta_i^{ve+} + \beta_i^{ve-}, \\
&\qquad g_i^e = \alpha_i^e * y_i + \beta_i^{ve+} - \beta_i^{ve-} + \beta_i^{pe+} - \beta_i^{pe-}, \\
&\qquad g_i^p = \alpha_i^p * y_i + \beta_i^{vp+} - \beta_i^{vp-} - \beta_i^{pe+} + \beta_i^{pe-}, \\
&\qquad \sum_{i=1}^{l} g_i^v = 0, \sum_{i=1}^{l} g_i^e = 0, \sum_{i=1}^{l} g_i^p = 0, \\
&\qquad 0 \leq \alpha_i^v \leq C^v, 0 \leq \alpha_i^e \leq C^e, 0 \leq \alpha_i^p \leq C^p, \\
&\qquad 0 \leq \beta_i^{ve+}, 0 \leq \beta_i^{ve-}, \beta_i^{ve+} + \beta_i^{ve-} \leq D^{ve}, \\
&\qquad 0 \leq \beta_i^{vp+}, 0 \leq \beta_i^{vp-}, \beta_i^{vp+} + \beta_i^{vp-} \leq D^{vp}, \\
&\qquad 0 \leq \beta_i^{pe+}, 0 \leq \beta_i^{pe-}, \beta_i^{pe+} + \beta_i^{pe-} \leq D^{pe} \\
&\qquad all\ for\ i = 1,\ ...,\ l
\end{aligned}
\tag{7}
$$

where β_i^{vp+}, β_i^{vp-}, β_i^{ve+}, β_i^{ve-}, β_i^{pe+}, β_i^{pe-}, α_i^v, α_i^e, α_i^p, λ_i^v, λ_i^e, and λ_i^p are all the Lagrangian multipliers. Among them, β_i^{vp+}, β_i^{vp-}, β_i^{ve+}, β_i^{ve-}, β_i^{pe+}, and β_i^{pe-} serve as bridges to relate the different classifiers. Then we use quadratic programming to solve the problem.

During the training process, the mapping functions interact through the similarity constraints. Since only the single feature space v is available during testing, we use the mapping function f_v to map the audio-visual feature to a real value, and use the decision function to determine the tag of the testing sample. Although only one feature space is used during testing, the information from other feature spaces of the training sample has remained in the trained model.

3.3 Extension to Incomplete Physiological Data

Since corrupt or missing data is frequent during physiological data collection, we extend the proposed emotion tagging method **so that it may use incomplete physiological data** during training. Multiple physiological signals are not usually corrupted simultaneously, so discarding all of the whole data instances containing invalid modalities results in a substantial amount of unusable data. **To fully utilize all available data**, we employ selection vectors during training to separately select the modalities that are not corrupted or missing and use **them** to train the emotion classifier from videos and the emotion classifiers from the selected physiological signals jointly, as shown in Eq. (8).

$$
\begin{aligned}
\underset{\Theta}{Min}\ & \frac{1}{2}\|w_v\|^2 + \frac{1}{2}\|w_e\|^2 + \frac{1}{2}\|w_p\|^2 + \\
& C^v \sum_{i=1}^{l} \xi_i^v + C^e \sum_{i=1}^{l} \mu_i^e \xi_i^e + C^p \sum_{i=1}^{l} \mu_i^p \xi_i^p + \\
& D^{ve} \sum_{i=1}^{l} \mu_i^e \eta_i^{ve} + D^{vp} \sum_{i=1}^{l} \mu_i^p \eta_i^{vp} + D^{pe} \sum_{i=1}^{l} \mu_i^e \mu_i^p \eta_i^{pe} \\
\textbf{s.t.}\ & \\
& \mu_i^e |f_v(v_i) - f_e(e_i)| \le \mu_i^e \eta_i^{ve} + \epsilon, \\
& \mu_i^p |f_v(v_i) - f_p(p_i)| \le \mu_i^p \eta_i^{vp} + \epsilon, \\
& \mu_i^e \mu_i^p |f_p(p_i) - f_e(e_i)| \le \mu_i^e \mu_i^p \eta_i^{pe} + \epsilon, \\
& y_i f_v(v_i) \ge 1 - \xi_i^v, \\
& \mu_i^e y_i f_e(e_i) \ge \mu_i^e (1 - \xi_i^e), \\
& \mu_i^p y_i f_p(p_i) \ge \mu_i^p (1 - \xi_i^p), \\
& \xi_i^v \ge 0,\ \xi_i^e \ge 0,\ \xi_i^p \ge 0, \\
& \eta_i^{vp} \ge 0,\ \eta_i^{ve} \ge 0,\ \eta_i^{pe} \ge 0 \\
& all\ for\ i = 1,\ ...,\ l
\end{aligned}
\tag{8}
$$

Compared to Eq. (6), Eq. (8) has two additional variables, i.e., μ^e and μ^p. They are selection vectors, indicating the availability of corresponding signals. Specifically, when the training samples have EEG features and the features of peripheral physiological signals, the corresponding μ_i^e and μ_i^p are equal to 1. Otherwise, they are set to 0.

This optimization problem can be translated to its dual problem by applying the Lagrange multiplier techniques. The dual problem is as follows:

$$
\begin{aligned}
\underset{\alpha}{Max}\ & -\frac{1}{2} \sum_{i,j=1}^{l} [g_i^v g_j^v K_v(x_i, x_j) + g_i^e g_j^e K_e(x_i, x_j) + \\
& g_i^p g_j^p K_p(x_i, x_j)] + \sum_{i=1}^{l} (\alpha_i^v + \mu_i^e \alpha_i^e + \mu_i^p \alpha_i^p) \\
\textbf{s.t.}\ & \\
& g_i^v = \alpha_i^v y_i - \mu_i^p (\beta_i^{vp+} - \beta_i^{vp-}) - \mu_i^e (\beta_i^{ve+} - \beta_i^{ve-}), \\
& g_i^e = \alpha_i^e y_i + \mu_i^e (\beta_i^{ve+} - \beta_i^{ve-}) + \mu_i^e \mu_i^p (\beta_i^{pe+} - \beta_i^{pe-}), \\
& g_i^p = \alpha_i^p y_i + \mu_i^p (\beta_i^{vp+} - \beta_i^{vp-}) - \mu_i^e \mu_i^p (\beta_i^{pe+} - \beta_i^{pe-}), \\
& \sum_{i=1}^{l} g_i^v = 0, \sum_{i=1}^{l} g_i^e = 0, \sum_{i=1}^{l} g_i^p = 0, \\
& 0 \le \alpha_i^v \le C^v, 0 \le \alpha_i^e \le C^e, 0 \le \alpha_i^p \le C^p, \\
& 0 \le \beta_i^{ve+}, 0 \le \beta_i^{ve-}, \beta_i^{ve+} + \beta_i^{ve-} \le D^{ve}, \\
& 0 \le \beta_i^{vp+}, 0 \le \beta_i^{vp-}, \beta_i^{vp+} + \beta_i^{vp-} \le D^{vp}, \\
& 0 \le \beta_i^{pe+}, 0 \le \beta_i^{pe-}, \beta_i^{pe+} + \beta_i^{pe-} \le D^{pe} \\
& all\ for\ i = 1,\ ...,\ l
\end{aligned}
\tag{9}
$$

From the above discussion, the proposed emotion tagging method without missing physiological data can be regarded as a special case of emotion tagging method with missing physiological data, where both μ^e and μ^p are 1 vector. The algorithm of our proposed model is outlined in Algorithm 1.

Algorithm 1 Algorithm of proposed model

```
Input: Audio-visual features $R^{|V|}$, EEG features $R^{|E|}$,
Peripheral features $R^{|P|}$, Emotional tag $Y$,
Weighted coefficients $\{C^v, C^e, C^p, D^{vp}, D^{ve}, D^{pe}\}$,
Selection vectors $\mu^e$ and $\mu^p$
Output: Predicted emotional tag
Choose the kernel function and project the input features into the kernel space;
Training phase
1.Solve the dual problem in Eq. (9) with quadratic programming;
2.Obtain the Lagrangian multiplier $\mu_i^p(\beta_i^{vp+} - \beta_i^{vp-})$, $\mu_i^e(\beta_i^{ve+} - \beta_i^{ve-})$, $\mu_i^p \mu_i^e(\beta_i^{pe+} - \beta_i^{pe-})$, $\alpha_i^v$, $\mu_i^e \alpha_i^e$, and $\mu_i^p \alpha_i^p$;
Testing phase
1.Estimate $f_v$ with the Lagrangian multiplier and projected audio-visual features;
2.Output the predicted emotional tag using Eq. (4);
```

3.4 Comparison to Related Work

Compared to SVM2K [21], which uses a similarity constraint of mappings from two feature spaces to combine kernel canonical correlation analysis (KCCA) and SVM into a single optimization problem, our model captures the relationships embedded in three modalities instead of just two views. In addition, SVM2K needs two views during testing, while our model requires only one modality. Our model can be seen as a way to learn a classifier using hidden information [22] or privileged information [23], since only one modality is available during testing but multiple modalities are used for training. The EEG features and features of peripheral physiological signals can be viewed as hidden information or privileged information that is used to help audio-visual features build a classifier. Unlike hidden information used as secondary features (proposed in [22]), which apply the ϵ-insensitive loss inequality constraints based on the assumption that secondary features are more informative for classification than the primary features, our model uses similarity constraints which are based on the assumption all feature spaces for classification are similarly useful. Our assumption is more general than Wang et al. 's [22]. The assumption of our method is also more general than SVM+ [23], which requires that privileged information and available information share the same slacking variable. Furthermore, both Wang et al.'s work [22] and SVM+ [23] only consider two modalities, not three modalities.

4 Experiments

4.1 Experimental Conditions

For these experiments, we evaluate our method on four benchmark databases: the DEAP database [7], the MAHNOB-HCI database [8], the USTC-ERVS database [9], and the LIRIS-ACCEDE database [10].

The DEAP database includes EEG signals and six kinds of peripheral physiological signals (EOG, EMG, GSR, RSP, TEMP, and BVP signals) from 32 participants as they were watching 40 stimulating music video clips. Two videos (experiment IDs: 17 and 18) cannot be downloaded from YouTube due to copyright issues. Therefore, we obtain 1216

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2702749, IEEE Transactions on Affective Computing

EEG and peripheral physiological segments corresponding to 38 stimulus videos. We take every physiological segment and its corresponding video as a sample. The self-assessment evaluations of users' induced emotions after video watching are reported in 9-point rating scales for valence and arousal.

The MAHNOB-HCI database includes EEG signals and four kinds of peripheral physiological signals (ECG, GSR, RSP, and TEMP signals) from 27 participants as they were watching 20 videos. Seven samples are removed because they don't have either a corresponding media file or gaze data. As a result, we obtain 533 EEG and peripheral physiological segments corresponding to 20 stimulus videos. We take every physiological segment and its corresponding video as a sample. Like the DEAP database, the emotional self-assessment evaluations are in 9-point rating scales for valence and arousal.

The USTC-ERVS database contains 197 EEG responses to 92 video stimuli from 28 users. We take every EEG response and its corresponding video as a sample. Users' emotional self-assessment evaluations consist of 5-point rating scales for both valence and arousal.

The LIRIS-ACCEDE database is composed of three sections: discrete, continuous, and MediaEval data sets. The continuous section contains the averaged GSR signal of 13 subjects recorded during they were watching 30 full-length movies. Each movie is annotated by 5 annotators along the induced valence axis, and 5 other annotators along the induced arousal axis. We cut these full movies into 60 seconds video clips since there were too few movies used to train a classifier. Thus there are 441 peripheral physiological segments corresponding to 441 video clips. We take every physiological segment and its corresponding video clip as a sample. Users' emotional self-assessment evaluations range from -1 to 1 for both valence and arousal for each frame. The averaged evaluations from 5 subjects are provided.

The four databases provide actual emotion labels of videos. The actual emotion is the affective response of a particular user to a video. It is context-dependent and subjective, and it may vary from one individual to another. Therefore, a video may have multiple ground truth evaluations, since it may be viewed by multiple subjects, with each subject providing different emotional self-assessments. In our emotion tagging approach, only audio-visual features are provided during testing, without any clues from subjects. Therefore, **we only consider aggregated emotion tags.** Furthermore, in the experiments, we adopt two categories, i.e., positive or negative valence and high or low arousal, instead of 5-point, 9-point or continuous rating scales, since the sample numbers are too small to conduct five or nine classification or regression on the databases. For example, there is only one video belonging to valence eight in the DEAP database. Binary classification of expected emotions is frequently used in the field of emotion tagging.

On the first three databases, the category label of a video is determined by the difference between its averaged ground truth evaluation and the averaged evaluations of all the videos, since a video is viewed by multiple subjects. We first average all the evaluations of a video and get its averaged ground truth evaluation. Then, we compare a video's averaged ground truth evaluation to the averaged evaluations of all the videos in the database. If the averaged ground truth evaluation of the video is larger than the averaged evaluations of all the videos, the video is regarded as positive valence or high arousal. Otherwise, the video is regarded as negative valence or low arousal. For the LIRIS-ACCEDE database, the category label of a video clip is determined by the average of the 60-second continuous annotations. If the averaged evaluation of the video is larger than 0, the video is regarded as positive valence or high arousal. Otherwise, the video is regarded as negative valence or low arousal. The sample size for valence and arousal on the four databases can be seen in Table 1.

We adopt the leave-one-video-out cross validation on the DEAP database, the MAHNOB-HCI database, and the USTC-ERVS database, and 10-fold cross validation on the LIRIS-ACCEDE database. Recognition accuracy, smaller F1-score, and unweighted average recall are used as performance metrics. Recognition accuracy measures overall classification accuracy without considering performance for each class. Smaller F1-score and unweighted average recall are used to ensure independence to unbalanced data distribution.

4.2 Experimental Results and Analysis with Complete Physiological Data

To evaluate our method with complete physiological data, we conduct five video tagging experiments, tagging videos using video content only, video content enhanced by EEG signals, video content enhanced by peripheral physiological signals, video content enhanced by both EEG and peripheral physiological signals through concatenating EEG features and peripheral physiological features as a feature vector, and the proposed method, which assigns emotion tags to videos from video content augmented by both EEG signals and peripheral physiological signals as two kinds of privileged information. We do not conduct experiments with peripheral physiological signals on the USTC-ERVS database, since it does not contain these signals of users. Likewise, experiments with EEG signals can not be conducted on the LIRIS-ACCEDE database, since it does not contain EEG signals of users.

Table 2 and Table 3 show video tagging results with complete physiological data for valence and arousal, respectively. From Table 2 and Table 3, we observe the following remarks:

1) Compared to video tagging from video content only, video tagging from video content enhanced by complete EEG signals achieves better performance for both valence and arousal. For valence and arousal tagging, our method increases recognition accuracy, F1-score, and average recall on three databases (the DEAP database, the MAHNOB-HCI database and the USTC-ERVS database) in most cases. This suggests that EEG signals which are only available during training are beneficial for building a better video emotion classifier.
2) Video tagging from video content enhanced by complete peripheral physiological signals outperforms the method using video content only for

TABLE 1
The sample size for valence and arousal on four databases

	valence				arousal			
	positive		negative		high		low	
	physiological segments	video clips	physiological segments	video clips	physiological segments	video clips	physiological segments	video clips
DEAP	544	17	672	21	800	25	416	13
MAHNOB-HCI	242	9	291	11	292	11	241	9
USTC-ERVS	64	30	133	62	162	70	35	22
LIRIS-ACCEDE	215	215	226	226	161	161	280	280

TABLE 2
Experimental results of video emotion tagging with complete physiological data for valence

		video only	with the help of EEG (CCA) [5]	with the help of EEG and peripheral (M-CCA)	with the help of EEG	with the help of peripheral	with the help of concatenated EEG and peripheral	with the help of EEG and peripheral (ours)
DEAP	accuracy	.658	.658	.737	.711	.737	.737	**.737**
	F1-score	.606	.606	.688	.686	.688	.706	**.722**
	recall	.655	.655	.736	.708	.736	.734	**.737**
MAHNOB-HCI	accuracy	.548	.600	.651	.752	.752	.754	**.797**
	F1-score	.470	.559	.592	.738	.738	**.741**	.713
	recall	.540	.597	.648	.773	.773	.774	**.865**
USTC-ERVS	accuracy	.868	.888	-	**.888**	-	-	-
	F1-score	.745	.814	-	**.817**	-	-	-
	recall	**.918**	.889	-	.884	-	-	-
LIRIS-ACCEDE	accuracy	.757	-	-	-	**.780**	-	-
	F1-score	.751	-	-	-	**.777**	-	-
	recall	.757	-	-	-	**.780**	-	-

TABLE 3
Experimental results of video emotion tagging with complete physiological data for arousal

		video only	with the help of EEG (CCA) [5]	with the help of EEG and peripheral (M-CCA)	with the help of EEG	with the help of peripheral	with the help of concatenated EEG and peripheral	with the help of EEG and peripheral (ours)
DEAP	accuracy	.737	.737	.816	.790	.790	.790	**.816**
	F1-score	.546	.615	.696	.692	.692	.714	**.741**
	recall	.713	.708	.811	.766	.766	.768	**.795**
MAHNOB-HCI	accuracy	.651	.700	.751	.850	.850	.850	**.852**
	F1-score	.632	.700	.738	.824	.824	.842	**.844**
	recall	.651	.707	.750	**.855**	**.855**	.850	.852
USTC-ERVS	accuracy	.797	**.822**	-	.797	-	-	-
	F1-score	.412	.407	-	**.524**	-	-	-
	recall	.648	**.684**	-	.681	-	-	-
LIRIS-ACCEDE	accuracy	**.696**	-	-	-	.687	-	-
	F1-score	.545	-	-	-	**.566**	-	-
	recall	**.669**	-	-	-	.662	-	-

both valence and arousal. For valence tagging, accuracy, F1-score, and average recall improve on all three databases (the DEAP database, the MAHNOB-HCI database, and the LIRIS-ACCEDE database). For arousal tagging, accuracy and average recall increase on two databases, and F1-score increases on all three databases. This further demonstrates that our proposed method successfully utilizes peripheral physiological signals only available during training to facilitate mapping between video content and emotion labels.

3) The performance of video tagging enhanced by complete EEG signals and video tagging enhanced by complete peripheral physiological signals is similar for both valence and arousal on the DEAP database and the MAHNOB-HCI database. This suggests that both EEG signals and peripheral physiological signals play helpful roles in assisting video tagging.
4) The performances of the proposed video tagging method and the video tagging method enhanced by concatenated EEG and peripheral physiological features are better than those of video tagging using video content only, video tagging from video content enhanced by EEG signals, and video tagging from video content enhanced by peripheral physiological signals, with higher accuracies, F1-scores, and average recalls in most cases. This demonstrates that EEG signals and peripheral physiological signals may be complementary in enhancing video emotion tagging.
5) The proposed video tagging method is superior to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2702749, IEEE Transactions on Affective Computing

the tagging method enhanced by concatenated EEG and peripheral physiological features. For valence tagging, accuracy and average recall improve on both databases, and F1-score improves on the DEAP database. For arousal tagging, accuracy and F1-score improve on both databases, average recall improves on the DEAP database. This demonstrates that the similarity constraints between classifiers during training can capture the intrinsic relationship between EEG and peripheral physiological signals more successfully compared to concatenating multiple features as one feature vector.

6) Compared to video tagging using video content only, the improvement of our proposed video tagging method is most significant on the MAHNOB-HCI database. It may be due to the more balanced data distribution in the valence and arousal space of the MAHNOB-HCI database.
7) The performance of arousal video tagging is better than that of valence tagging on the DEAP database and the MAHNOB-HCI database, indicating that arousal recognition may be easier than valence recognition. The data distribution in the arousal space is more balanced than that in the valence space for these two databases. Valence recognition is more accurate than arousal recognition on the USTC-ERVS database and the LIRIS-ACCEDE database because the data in the valence spaces are more balanced.

We also adopt the 5X2 cross validation paired t-test [24] to check whether the improvement of the proposed method compared to the other methods is significant or not based on the F1-scores on the DEAP database and the MAHNOB-HCI database. The p-values are all less than 0.05 when the F1-scores obtained by our method are larger than those from other methods. This demonstrates that the improvement is significant.

4.3 Comparison to Other Methods

Currently, little work in the field of video emotion tagging exploits the relationship between users' emotional responses and stimuli. As mentioned in Section 2, there are only two works using implicit hybrid video emotion tagging [5] [14]. Since the database used by Han et al. [14] is not publicly available, we cannot compare our work with theirs. Here, we compare our work with Wang et al.'s.

Wang et al. [5] proposed to tag videos' emotions with the aid of EEG signals by constructing a new video feature space that exploits the relationship of EEG signals and video content using CCA. Like us, they verify their proposed method on the DEAP database, the MAHNOB-HCI database, and the USTC-ERVS database. They also adopted two categories, i.e., positive or negative valence and high or low arousal, instead of 5- or 9-point rating scales. However, their strategy to change users' self-reported 5- or 9-point rating scales to binary emotion labels is different from ours. They assign binary labels to a video by comparing the ground truth evaluation of a video to the middle value of evaluations. Thus, a video may have different binary labels since it may be evaluated by multiple subjects. Since only video content is used during testing, a single unique label for a video is more reasonable. Although we could not compare our experimental results directly with theirs due to the label difference, we replicate their experiments using their proposed method. Rather than integrating multiple physiological signals to facilitate video emotion tagging as we did, Wang et al. [5] enhanced emotion tagging with only one kind of physiological signal (EEG). We extend their proposed implicit tagging with multi-set CCA (M-CCA), which optimizes an objective function of the correlation matrix of the canonical variate from multiple random vectors such that the canonical variate achieves maximum overall correlation. This allows us to more accurately compare our methods. The experimental results of this comparison are shown in Table 2 and Table 3.

From Table 2 and Table 3, we find that our method for video tagging enhanced by multiple physiological signals outperforms Wang et al.'s method of video tagging enhanced by one physiological signal. This further proves that multiple physiological signals are complementary in enhancing video emotion tagging. In most cases, recognition accuracy, F1-score, and average recall are higher for our method than that for Wang et al.'s method. This indicates that our method is more effective at taking advantage of physiological signals for video emotion tagging during training, since its objective function is to minimize tagging error directly, while the goal of Wang et al.'s is to construct a new video feature space processing the highest Pearson correlation coefficients with the physiological feature space.

4.4 Experimental Results and Analysis with Incomplete Physiological Data

To evaluate our method's performance when physiological data is incomplete, we conduct four video tagging experiments: one using video content enhanced by incomplete EEG signals, one using video content enhanced by incomplete peripheral physiological signals, one using video content enhanced by incomplete EEG and peripheral physiological signals through concatenating EEG and peripheral physiological features as a feature vector, and one using video content enhanced by both incomplete EEG signals and peripheral physiological signals as two kinds of privileged information. We randomly miss 5%, 10%, 15%, 20%, or 25% of EEG features and peripheral physiological features. We conduct experiments on the DEAP database and the MAHNOB-HCI database, since they contain both EEG signals and peripheral physiological signals. We conduct twenty times experiments for each missing rate and select the average accuracy, F1-score, and average recall of the middle ten times.

Fig.2 shows video tagging results with incomplete physiological data for valence and arousal on the DEAP database and the MAHNOB-HCI database. From this figure, we can obtain the following remarks:

1) For all video tagging experiments, recognition accuracy, F1-score, and average recall decrease continuously on both databases as loss of EEG signals and peripheral physiological signals increases. This is reasonable since fewer EEG signals and peripheral physiological signals provide less contributions to

video emotion classifiers from video content, and thus results in decreased performance.

2) Compared to Table 2 and Table 3, we find that the experimental results with incomplete physiological data are superior to the experiments which use video signals only and inferior to the experiments which use video signals with the help of complete physiological signals in most cases. This further demonstrates that EEG signals and peripheral physiological signals are helpful to video tagging.
3) When the four video tagging experiments have the same percentage of incomplete physiological data, the performance of the proposed video tagging method and the video tagging method enhanced by concatenated EEG and peripheral physiological features is better than that of video tagging from video content enhanced by EEG signals and video tagging from video content enhanced by peripheral physiological signals, with higher accuracies, F1-scores, and average recalls for both valence and arousal tagging on the two databases. This demonstrates that EEG signals and peripheral physiological signals may be complementary in enhancing video emotion tagging.
4) When the four video tagging experiments have the same percentage of incomplete physiological data, the proposed video tagging method achieves better results than the video tagging method enhanced by concatenated EEG and peripheral physiological features in most cases. For valence tagging, the proposed method has the highest recognition accuracy, F1-score, and average recall on both databases. For arousal tagging, the proposed method has the highest recognition accuracy and average recall on both databases. This demonstrates that the proposed method is more powerful in capturing the relationship between EEG and peripheral physiological signals.
5) When the four video tagging experiments have the same percentage of incomplete physiological data, we observe that video tagging enhanced by incomplete EEG signals has better performance than video tagging enhanced by incomplete peripheral physiological signals in the valence space, while video tagging enhanced by incomplete peripheral physiological signals has better performance than video tagging enhanced by incomplete EEG signals in the arousal space on the both databases. This shows EEG signals are more helpful to video tagging in the valence space and peripheral physiological signals are more helpful to video tagging in the arousal space. This is consistent with our all observations.
6) For all video tagging experiments, the accuracy differences, F1-score differences, and average recall differences between video tagging enhanced by 95% physiological signals and video tagging enhanced by 75% physiological signals are larger on the MAHNOB-HCI database compared the DEAP database. This may be due to the more balanced data distribution on the MAHNOB-HCI database.
7) For all video tagging experiments, the accuracy differences, F1-score differences, and average recall differences between video tagging enhanced by 95% physiological signals and video tagging enhanced by 75% physiological signals are larger in the arousal space than in the valence space on both databases. The reason may be that the data distribution in the arousal space is more balanced than that in the valence space for these two databases.

5 Conclusions

In this paper, we propose an implicit hybrid video emotion tagging approach that integrates video content and users' multiple physiological responses, which are only required during training. Specifically, we propose similarity constraints on the emotion classifiers from videos and the emotion classifiers from available physiological signals to capture the nature of the relationships among users' physiological responses, video content, and emotion labels. The experimental results on four benchmark databases demonstrate that our proposed method with the help of physiological signals outperforms the baseline method, which uses video signals only. Multiple physiological signals are complementary in enhancing video tagging. For the experiments with complete physiological data, our proposed method using EEG signals and peripheral physiological signals has the best performance. For the experiments with incomplete physiological data, our proposed method also shows that both EEG signals and peripheral physiological signals play helpful roles in video tagging. Furthermore, the comparison to related work shows our approach is superior to current work, as it explicitly captures the embedded relationships among multiple modalities and emotion labels.

Acknowledgments

This work has been supported by the National Science Foundation of China (grant nos. 61473270, 61228304, and 61175037) and the project from Anhui Science and Technology Agency (1508085SMF223).

References

[1] S. Wang and Q. Ji, "Video affective content analysis: a survey of state of the art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[2] A. Vinciarelli, N. Suditu, and M. Pantic, "Implicit human-centered tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 1428–1431, 2009.

[3] M. Soleymani, "Implicit and automated emotional tagging of videos," *Emotion*, 2011.

[4] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users eeg and video content," *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1257–1283, 2014.

[5] S. Wang, Y. Zhu, L. Yue, and Q. Ji, "Emotion recognition with the help of privileged information," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 1–1, 2015.

[6] S. Chen, S. Wang, C. Wu, Z. Gao, X. Shi, and Q. Ji, "Implicit hybrid video emotion tagging by integrating video content and users multiple physiological responses," in *23rd International Conference on Pattern Recognition in Cancun, Mexico*, 2016.

[7] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[8] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[9] S. Wang, Z. Liu, S. Lv, and Y. Lv, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[10] T. Li, Y. Baveye, C. Chamaret, and L. Chen, "Continuous arousal self-assessments validation using real-time physiological responses," in *International Workshop on Affect and Sentiment in Multimedia*, 2015, pp. 39–44.

[11] F. Nack, C. Dorai, and S. Venkatesh, "Computational media aesthetics: Finding meaning beautiful," *IEEE MultiMedia*, vol. 8, no. 4, pp. 10–12, 2001.

[12] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[13] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2012, pp. 3304–3309.

[14] J. Han, X. Ji, X. Hu, L. Guo, and T. Liu, "Arousal recognition using audio-visual features and fmri-based brain response," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 1–1, 1949.

[15] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[16] S. Koelstra, C. Muhl, and I. Patras, "Eeg analysis for implicit tagging of video data," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 104–105.

[17] S. Koelstra, A. Yazdani, M. Soleymani, C. Mhl, J. S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos," in *Brain Informatics, Ser Lecture Notes in Computer Science*, 2010, pp. 89–100.

[18] J. Kim and E. Andr, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–83, 2008.

[19] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01)*, vol. 1. IEEE, 2001, pp. 73–76.

[20] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, 2006.

[21] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-taylor, and S. Szedmak, "Two view learning: Svm-2k, theory and practice," in *Advances in neural information processing systems*, 2005, pp. 355–362.

[22] Z. Wang and Q. Ji, "Classifier learning with hidden information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4969–4977.

[23] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 22, no. 5-6, pp. 544–557, 2009.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

Shangfei Wang received her BS in Electronic Engineering from Anhui University, Hefei, Anhui, China, in 1996. She received her MS in circuits and systems, and the PhD in signal and information processing from University of Science and Technology of China (USTC), Hefei, Anhui, China, in 1999 and 2002. From 2004 to 2005, she was a postdoctoral research fellow in Kyushu University, Japan. Between 2011 and 2012, Dr. Wang was a visiting scholar at Rensselaer Polytechnic Institute in Troy, NY, USA. She is currently an Associate Professor of School of Computer Science and Technology, USTC. Dr. Wang is an IEEE and ACM member. Her research interests cover computation intelligence, affective computing, and probabilistic graphical models. She has authored or co-authored over 70 publications.

Shiyu Chen received her BS in computer science from Anhui University in 2015, and she is currently pursuing her MS in Computer Science in the University of Science and Technology of China, Hefei, China. Her research interesting is affective computing.

Qiang Ji received his PhD in Electrical Engineering from the University of Washington. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He recently served as a program director at the National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, the Dept. of Computer Science at University of Nevada at Reno, and the US Air Force Research Laboratory. Prof. Ji currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI.

Prof. Ji's research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published over 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including NSF, NIH, DARPA, ONR, ARO, and AFOSR as well as by major companies including Honda and Boeing. Prof. Ji is an editor on several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. Prof. Ji is a fellow of IAPR and IEEE.

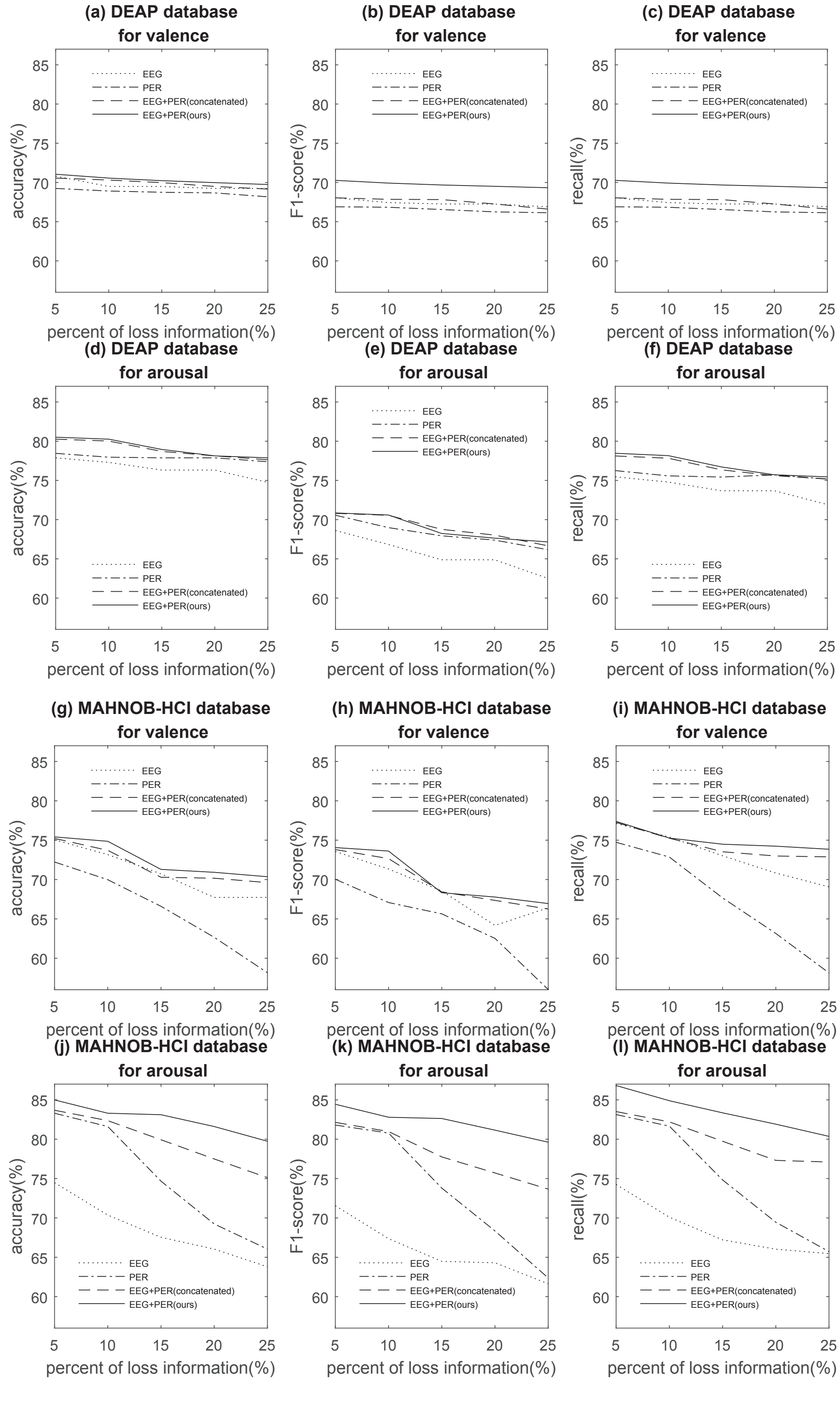

Fig. 2. The experimental results of video emotion tagging with incomplete physiological data on the DEAP database and the MAHNOB-HCI database, 'EEG' indicates video tagging from video content enhanced by incomplete EEG signals; 'PER' refers to video tagging from video content enhanced by incomplete peripheral physiological signals; 'EEG+PER(concatenated)' means video tagging from video content enhanced by incomplete EEG and peripheral physiological signals through concatenating EEG features and peripheral physiological features as a feature vector; and 'EEG+PER(ours)' means video tagging from video content enhanced by both incomplete EEG signals and peripheral physiological signals as two kinds of privileged information. (a), (b), and (c) respectively show accuracy, F1-score, and average recall on the DEAP database for valence tagging; (d), (e), and (f) respectively show accuracy, F1-score, and average recall on the DEAP database for arousal tagging; (g), (h), and (i) respectively show accuracy, F1-score, and average recall on the MAHNOB-HCI database for valence tagging; and (j), (k), and (l) respectively show accuracy, F1-score, and average recall on the MAHNOB-HCI database for arousal tagging.