IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 15, 2018, accepted September 24, 2018, date of publication October 9, 2018, date of current version November 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2874892

Towards Generic Modelling of Viewer Interest Using Facial Expression and Heart Rate Features

PRITHWI RAJ CHAKRABORTY[1],
DIAN WIRAWAN TJONDRONEGORO[1], (Senior Member, IEEE),
LIGANG ZHANG[2], AND VINOD CHANDRAN[3], (Senior Member, IEEE)
[1]IT Discipline, School of Business and Tourism, Southern Cross University, Bilinga, QLD 4225, Australia
[2]School of Engineering and Technology, Central Queensland University, Brisbane, QLD 4000, Australia
[3]School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4001, Australia

Corresponding author: Prithwi Raj Chakraborty (prithwi.chakraborty@scu.edu.au)

This work was supported by the Queensland University of Technology under Post-Graduate Research Award and Write-Up Scholarships.

ABSTRACT Automatic detection of viewer interest while watching video contents can enable multimedia applications, such as online video streaming, to recommend contents in real time. However, there is yet a generic model for detecting viewer interest that is independent of subject and content while using non-invasive sensors in near-natural settings. This paper is the first attempt at solving this issue by investigating the feasibility of a generic model for detecting viewer interest based on facial expression and heart rate features. The proposed model adopts deep learning features, which are trained and tested using multi-subjects' data across different video stimuli domains. The experimental results show that the generic model can reach a similar accuracy to a domain-specific model.

INDEX TERMS Facial expression, heart rate, heart rate variability, viewer interest.

I. INTRODUCTION

Interest is a distinct and anticipatory type of emotion that characterizes people's experience with new objects, events, or situations [1], [2]. Automatic detection of interest can be used for user-centric multimedia applications, including implicit video/image tagging, audience engagement measurement, and automation of video recommendation tools [24]. Viewer interest is defined as a type of interest that occurs during video viewing [3], which encompasses a certain level of engagement, anticipated effort, and attentional activity [4], [5]. Viewer interest is detectable through video observation and physiological signals, including facial expression, eye gaze, heart rate, and skin response. However, these signals are not always visible, accessible, distinctive and synchronized. The signals are also prone to subjective bias and individual preference towards the content type, event duration, and viewing conditions.

Studying viewer interest in each of the different domains (or genres) of video stimuli presents unique challenges due to difference in genre, style, and aesthetics [14]. Short movie clips are currently the most commonly used type of stimuli, as movies can generally evoke interests from a wide-range of viewers. For example, compared to movie news video would be more subjected to biases due to topical interest, political standpoint, and socio-economic backgrounds. However, compared to sports videos, which have a more structured highlights, selecting the most suitable stimuli segments from movie videos generally requires a deeper understanding of the whole movie's story flow and narrative.

Training new model for detecting viewer interest for each of the different video domains and subjects would be impractical in real world applications. Previous work have shown that viewer interest models trained for a particular sports type still outperformed those trained with multiple sports types [6]. There is no previous study that has established a generic model trained with multi-subjects' data across different video stimuli domains. This paper is the first attempt at investigating the suitability of a generic model based on facial expression and heart rate features for detecting viewer interest. The experiments, which incorporated *domain-general*, *cross-domain*, and *domain-specific* training and testing approaches using sports and movie video stimuli, demonstrate novel insights on how data homogeneity can influence the detection performance.

The key findings of this paper are:

(1) It is feasible to achieve a generic (domain-general) model for viewer interest detection using deep learning approach that can reach a similar accuracy to the

domain-specific models, which means that we can potentially avoid hand-crafting features for future studies.

(2) Heart rate features are found to be the better indicators for viewer interest than facial expression, and the data collection can be less intrusive during real world implementations, due to privacy concerns from recording facial data.

II. RELATED WORK

The choice of multimedia stimuli and features is crucial for developing a machine learning model for detecting viewer interest. Existing studies have used static stimuli including random images of geometrical shapes, random pictures from books and journals, classical paintings [7]–[10], visual poetry, mannequins, live person faces, and lifeless objects [2], [11]. Some studies directly asked explicit indication of interests using manual approach without sensing [2]. Other studies used content-based features to predict the specific contents that can potentially evoke interests. For example, image-based raw pixel value, histogram, self-similarity and global texture distribution features were used for measuring interestingness of a sequence of images [12]. Similarly, auditory energy and color histograms from the video frame are used with machine learning techniques to predict potential level of affect and interest [13]. A number of low-to-mid level audio-visual and temporal features were used to predict aesthetic, affect and interest level from movie clips compared to viewer ratings [14].

Asking for explicit indication of viewer interest can be too obtrusive during data collection. Therefore, studies have started to propose methods to measure interestingness based on indirect cues that can be sensed. Mouse cursor activity has been used to measure interests in online video. Interest score was estimated based on mouse movement features within the smaller and longer time spans [15]. Intriguingly, the choice of interesting contents by infants was found to be statistically correlated with heart rate, visual fixation, and face activity [11]. Likewise, head-eye responses and facial expressions can help to predict viewer interest based on fuzzy-logic-based fusion of binary and probabilistic measures [16]. More recently, student's engagement levels during online writing can be predicted using local binary pattern features from Kinect-based facial recording and heart rate responses [17].

Based on the above-mentioned studies, facial expression and physiological responses are the two most extensively used viewer responses for detecting interest and engagement levels. Facial expression analysis methods generally classify emotion states using appearance, geometric and motion-based features extracted from different parts of faces including head, eye (eyelid, eyebrow), mouth (nose, lip), and cheek regions [17]–[21]. Study based on facial landmark movements has found that the upper-part of human face can better detect interest-evoking highlights more than the lower part [22]. Recent work has also used aggregated visual features, such as LBP, local-LBP, SIFT and histogram extracted from face, upper body, and background scene to classify group-level emotion within social group images [23]. Gabor filter features were used to classify facial expressions of participants as a way to obtain implicit tagging of image sequence achieving 51-60% accuracy [24].

As a type of emotion, interests are generally more subtle, therefore their recognition is more difficult [25]. Facial activity sometimes do not show any change on emotional feelings, and subjects may not express their interests through facial expression [25]. Physiological responses are considered to be good indicators of emotional states. For example, affective indexing of music excerpts can be achieved based on a linear combination of decisions with equal-weights from EEG, ECG, GSR, and head pose signals [26]. An EEG-based study achieved a mean accuracy of about 64% for arousal rating and 56% for valence ratings of viewers in response to images and movie clips [27]. Viewers' fMRI response were used along with content-based features to measure arousal level of video clips with a 92-93% accuracy [28].

Heart rate has been found complementary to facial expression for viewer interest detection [29] and directly proportional to interest and pleasantness, but inversely proportional to our cognitive anticipation to stimuli [30], [31]. More specifically, acceleration and deceleration of heart rate correlate with engagement and negative valence cues (e.g., visual fixation) [11]. It also indicates short-term attention [32]. Heart rate variability (HRV) features is inversely correlated with fear, sadness, and happiness [33]–[35]. Other heart rate features to measure interests include acceleration, deceleration, beats-per-minute readings, intervals: in-between heart beats, and energy in frequency bands [11], [36]. Fast and significant changes in these features usually indicate emotion-evoking events, which vary across different stimuli types [37]–[39]. Statistical features including standard deviation, skewness, and kurtosis of heart rate were used in measuring emotions in an arousal-valence space [40]. The beat-to-beat intervals (better known as RR intervals) are direct measure of HRV and considered as reliable markers of human's affective responses [31]. Energy in different frequency bands represents sympathetic modulation excitement such as: low frequency (LF), range from 0.04 to 0.15 Hz; high frequency (HF), range from 0.15 to 0.40 Hz; and very low frequency (VLF), range from 0.0033 to 0.04 Hz [31]. Negative emotions including fear, anxiety, and pain decrease the HF, while anger highly correlates with the ratio between LF and HF [41]. Statistical HRV features, such as standard deviation of the mean beat-to-beat intervals along with VLF, has been shown to have a significant correlation with depression [42], [43].

Deep learning techniques have been found effective for emotion recognition [44]. A common approach is to use a stack of neural network layers to automatically extract low-level features in lower layers, and high-level features in subsequent layers [45], thereby eliminating the need for manual engineering of features. State-of-the-art results have been shown for applying deep learning for object recognition, object classification and learning semantic visual

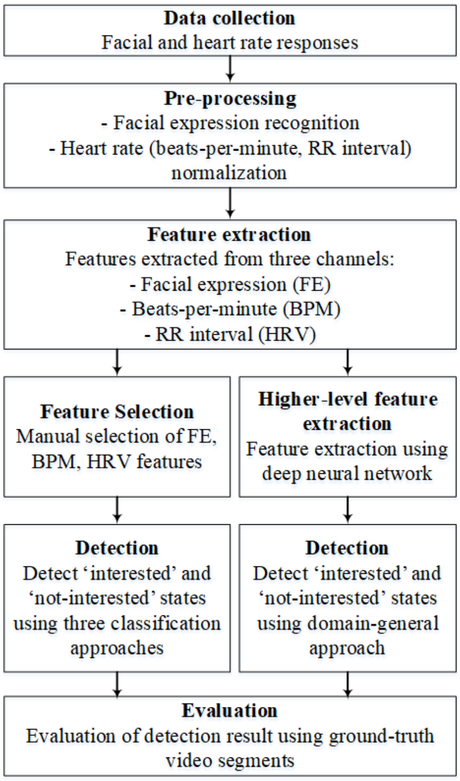

Data collection
Facial and heart rate responses
Pre-processing
- Facial expression recognition
- Heart rate (beats-per-minute, RR interval) normalization
Feature extraction
Features extracted from three channels:
- Facial expression (FE)
- Beats-per-minute (BPM)
- RR interval (HRV)
Feature Selection
Manual selection of FE, BPM, HRV features
Higher-level feature extraction
Feature extraction using deep neural network
Detection
Detect 'interested' and 'not-interested' states using three classification approaches
Detection
Detect 'interested' and 'not-interested' states using domain-general approach
Evaluation
Evaluation of detection result using ground-truth video segments

FIGURE 1. Experimental protocol.

features [46], however, it has not been established for viewer interest detection.

III. EXPERIMENT PROTOCOL

The overall experimental protocol is illustrated in Fig. 1. Facial and heart rate responses were collected from participants in three separate user studies using soccer, tennis, and movie stimuli respectively. The recorded facial videos were processed to extract intensity scores of facial expression categories. Heart rate data were recorded as beats-per-minute readings and RR intervals. Data from these three channels were pre-processed for normalization and handling missing values. From these channels, a total of 17 features were extracted (discussed in Section IV) and three classification approaches were benchmarked for their robustness in detecting the 'interested' states. Full details on the user studies and some parts of the analysis can be found in [47]. This paper will summarize the experiments, while focusing on additional heart rate and deep learning features, and extended classification methods, which produce the new results.

A. PARTICIPANTS

Subjects were between 21 and 30 years old (mean = 26, standard deviation = 3), recruited from university student/staff with consents. A preliminary screening ensured that the subjects were not familiar with the stimuli clips and none of them had eye or heart condition. Data was collected from three separate user studies. The first two studies with sports stimuli were attended by a total of 12 subjects (11 male, 1 female), among which 9 subjects participated in both. The last study with movie stimuli was attended by a total of 20 subjects (16 male, 4 female), among which 7 subjects have also participated in the studies with sports stimuli.

B. STIMULI PREPARATION

The sports user study used a total of 3 soccer and 2 tennis video clips. The duration of these clips varied between 9 and 20 minutes. The source videos were selected from different international leagues to maintain data heterogeneity. The soccer clips consisted of 3 *goal*, 14 *shot-on-goal*, and 5 *foul* events. The tennis clips used in the study included 29 *rally* events, each containing a sequence of back and forth shots between players within a point. Further details of the sports stimuli and the selection procedure can be found in [6].

The movie user study used 20 movie clips, collected from MAHNOB-HCI (8 clips) [48], LIRIS-ACCEDE (5 clips) [49], and YouTube (7 clips). The durations of the movie clips were between 25 and 112 seconds (average of 68.25 ± 43.66 seconds), which were long enough to evoke viewers' emotions [49]. Both manual and automatic processing were applied for selecting the movie clips, as described in the following sub-sections. During stimuli selection, affective ratings of the clips were considered rather than their genres. Clips which are represented as emotionally 'neutral' were not considered, since the goal was to obtain clips that are potentially able to evoke viewers' 'interest'.

1) MOVIE STIMULI SELECTION FROM MAHNOB-HCI

The dataset contains 20 movie clips and two types of subjective feedbacks for them. The feedbacks include emotion tags and valence-arousal scores (in 9-point scale) collected from 27 participants for each clip, both of which were used to select the stimuli clips (detail procedure is discussed in the following steps). Valence and arousal scores are dimensional scores indicating the level of perceived pleasantness and activation. A two-step procedure was followed to select the clips from MAHNOB-HCI dataset. To ensure the robustness of the clip selection, the procedure considered both categorical and dimensional feedbacks. Firstly, the procedure pre-selected a number of clips using dominant emotion tags computed with histogram plot (Step 1). Then the pre-selection was filtered by a valence-arousal plot (Step 2).

Step 1: A histogram plot (Fig. 2) for each movie clip was drawn using the emotion tags of the dataset. Each such plot computed how many times each of the 13 emotion tags occurred (i.e., rated by participants) for a particular clip. For each clip, a 'dominant emotion' tag was computed based on the emotion that was tagged for the maximum times for that clip. For example, clip15 was determined as a 'neutral' clip as it received 'neutral' tag for the maximum number of times. Movie clips with (dominant) emotion tags including happiness, sadness, fear, and disgust, were primarily selected.

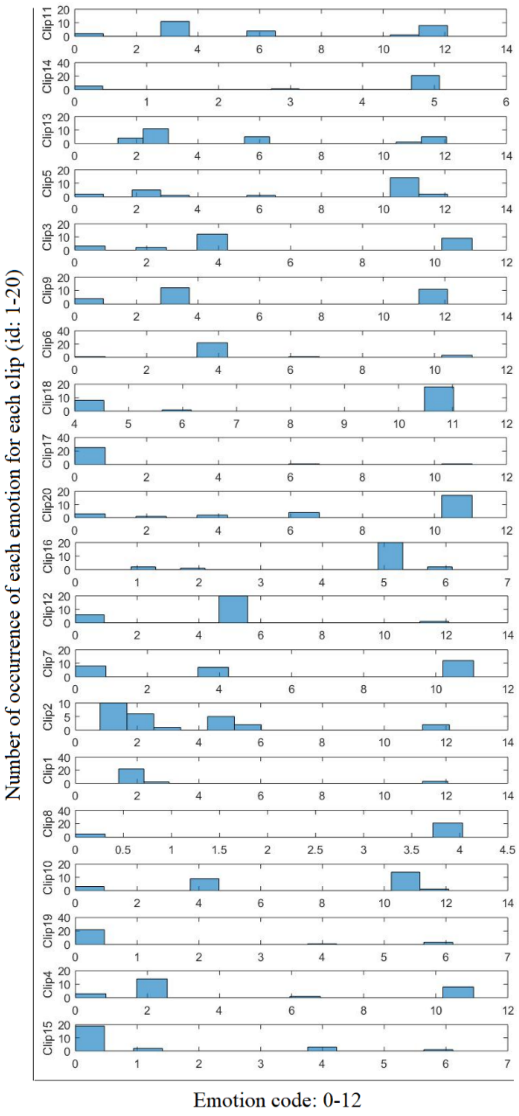


FIGURE 2. Histogram plots of 20 clips from the MAHNOB-HCI dataset. Y axis stands for the number of occurrences of each emotion tag and clip id (1 - 20), X axis stands for the 13 emotion tag id (id: 0-12).

Step 2: An ellipse was plotted for each of the 20 clips into an arousal vs. valence space, using the means and standard deviations of the valence and arousal scores for that particular clip. The center and radius of each ellipse were denoted by the mean and standard deviation of the valence and arousal scores (Fig. 3, reproduced from [50, Fig. 3]). The selected clips in Step 1 were furthered filtered using ellipse's positions and sizes. Among the movie clips those were primarily selected in Step 1, a total of 8 clips with the maximum valence and arousal scores tagged with those emotions were selected.

2) MOVIE STIMULI SELECTION FROM LIRIS-ACCEDE

The dataset contains 9,800 short clips, with lengths between 8 and 12 seconds. LIRIS-ACCEDE dataset includes regressed valence-arousal values, which were reported using a 5-point manikin. A 5-component k-mean method is used to compute the clusters for these regressed valence-arousal values (depicted in Fig. 4 plotted in a valence vs arousal space). The extreme (comparatively higher) values along the X (valence) and Y (arousal) axes were used to select a total of 5 clips. Mean of the valence-arousal values were used to approximate categorical emotion tags for these 5 clips. The tags were confirmed by manually viewing the clips.

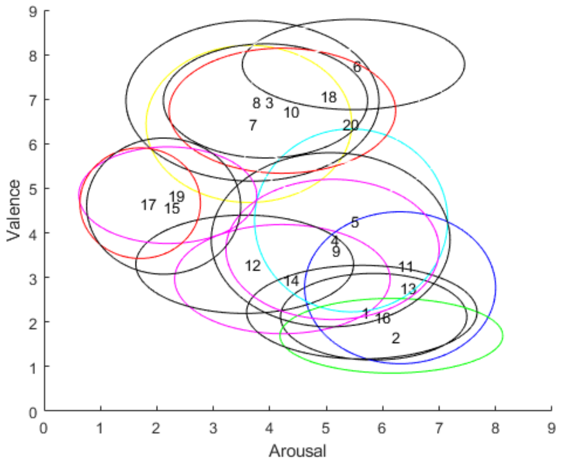


FIGURE 3. Arousal and valence scores from MAHNOB-HCI dataset plotted with ellipse. Means and standard deviations of the arousal-valence scores are used as centers and radiuses of the ellipses. The clip ids (1-20) are marked at the center of each clip.

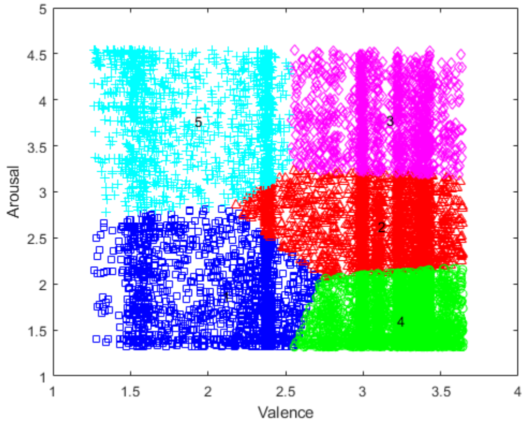


FIGURE 4. The 5-component (K = 5) K-mean cluster plots of the regressed valence-arousal values (obtained from LIRIS-ACCEDE dataset), where each color represents a separate cluster (id: 1-5).

3) MOVIE STIMULI SELECTION FROM YOUTUBE

7 clips were manually obtained from YouTube by searching for movie clips (surprise, happy, and fear categories) with keywords such as 'the best horror scene' and 'the best comedy scene'. The selection was confirmed by visually observing each clip. A full list of the 20 selected clips from is presented in Table 1.

C. DATA COLLECTION PROCEDURE

For collecting viewer interest data, facial expression data was recorded using a Panasonic full HD camcorder at

TABLE 1. Details of source and duration of 20 movie stimuli used.

Original Clip	Source	Duration (min)	Emotion Tag
80.avi	MAHNOB-HCI	1:37	Happiness
ACCEDE00317.mp4	LIRIS-ACCEDE	0:08	Happiness
58.avi	MAHNOB-HCI	0:59	Happiness
ACCEDE00362.mp4	LIRIS-ACCEDE	0:10	Happiness
The Kid	YouTube	2:19	Happiness
The Great Dictator	YouTube	1:48	Happiness
30.avi	MAHNOB-HCI	1:11	Fear
ACCEDE00911.mp4	LIRIS-ACCEDE	0:09	Fear
107.avi	MAHNOB-HCI	0:35	Fear
The Visit	YouTube	1:05	Fear
Insidious	YouTube	2:14	Fear
55.avi	MAHNOB-HCI	1:17	Anger/ Fear
ACCEDE00997.mp4	LIRIS-ACCEDE	0:11	Anger
ACCEDE00912.mp4	LIRIS-ACCEDE	0:08	Anger
111.avi	MAHNOB-HCI	1:54	Sadness
138.avi	MAHNOB-HCI	1:57	Sadness
Interestelar	YouTube	1:25	Surprise
Disturbia	YouTube	1:07	Surprise
Stand by Me	YouTube	1:41	Disgust
69.avi	MAHNOB-HCI	0:59	Disgust

a 25 frame-per-second (fps) sampling rate. Heart rate data was collected using Mio Alpha (wrist-worn) and Polar H7 (chest-strapped) devices at 1/3 Hz and 1 Hz sampling rates respectively. Polar H7 provided both beats-per-minute and RR interval data, while Mio Alpha only gave beats-per-minute, and the resolution of the RR interval data varied between 600 and 900 milliseconds. These two heart rate sensors were used to record redundant data as backup and complement one another.

The stimuli were presented using a 22-inch high-definition monitor in a closed room environment. The sports stimuli had a frame resolution of 1280 × 720 pixels and a frame rate of 25 fps. The resolution of the movie stimuli varied between 640 × 386 and 1280 × 800 as they were collected from two public datasets and YouTube. A viewing distance of 80-90 cm between the subject and the monitor was kept during the study.

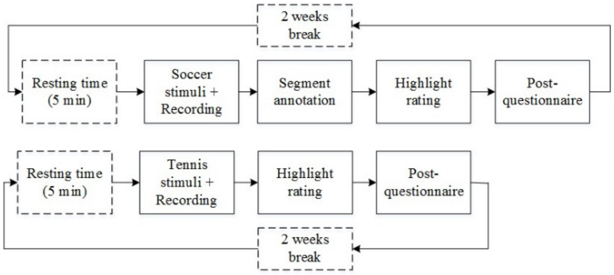


FIGURE 5. Data collection protocol for soccer and tennis clips.

1) DATA COLLECTION USING SPORTS STIMULI

User study with soccer and tennis stimuli was conducted in identical but separate sessions. Each session involved three steps – recording, feedback, and questionnaire as shown in Fig. 5. The simultaneous viewing and recording of data commenced after a short resting time to stabilize heart rate. After each clip viewed, the subject annotated ''potentially interesting'' segments, and identified a number of soccer highlight segments from the viewed clip. Then, each subject went through a questionnaire to rate about general information in the viewed clip.

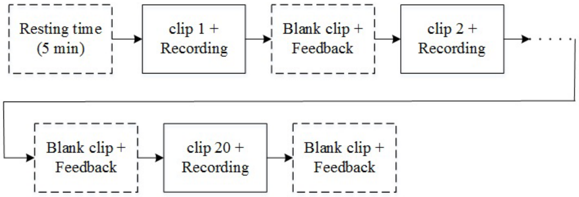


FIGURE 6. Data collection protocol for movie clips.

2) DATA COLLECTION USING MOVIE STIMULI

User study with movie stimuli started with a 5-minutes resting period after the subject completed a briefing session that explained the procedures, interfaces, setups, and annotation methods. Fig. 6 shows the details of the experiment protocol. The movie clips were sorted in a happy-fear-sadness-anger-disgust order (as shown in Table 1). The clips were sorted in a low-to-high intensity order if there were more than one clip for a particular emotion type. The intensity was measured by manually viewing the clips. The primary goal for sorting clips was to see if the sorting has any influence on how the participant respond to the stimuli (findings are discussed in Section IV). A 2-minute blank clip was inserted between two consecutive movie clips, so that it neutralizes the subject's responses (recording was paused during this time) and provide time for completing feedbacks using a paper-based survey form. The feedbacks included: binary rating (1 - interesting, 0 - not interesting); associated emotion (neutral, happiness, fear, sadness, anger, surprise, disgust); valence (1: unpleasant to 5: pleasant); and arousal scores (1: calm to 5: active).

D. GROUND TRUTH

MAHNOB-HCI and LIRIS-ACCEDE datasets do not provide viewer interest annotations. Therefore, this study collected subjects' annotations and ratings during the data collection. Each sports and movie video has received a binary rating ('interesting' or not) for each clip. Ground truth was determined from subjects' annotations and ratings in response to soccer, tennis, and movie clips. The ground truth was prepared with the starting and ending time indices of segments those receive more than 50% subject agreements (i.e. majority vote).

IV. FEATURE EXTRACTION

The features used in this study are summarized in Table 2, which will be described in this section. Facial expression and heart rate responses are used to extract features from three channels – facial expression (FE), beats-per-minute

TABLE 2. Features used in classification of viewer interest.

Channel	Features
FE	Emotion intensity scores for positive, negative, and neutral categories, (one) deep learning feature
BPM	Beats-per-minute readings, derivative, variance, range, HF, LF, and LF/HF, (four) deep learning features
HRV	RR intervals, NN50, RMSSD, SDNN, HF, LF, and LF/HF, (three) deep learning features

readings (BPM), and RR intervals as heart rate variability (HRV). Features from each channel are used separately to model viewer interest. During feature selection, features are hand-picked based on a manual evaluation that checks for redundancy (correlation) and high number of missing instances.

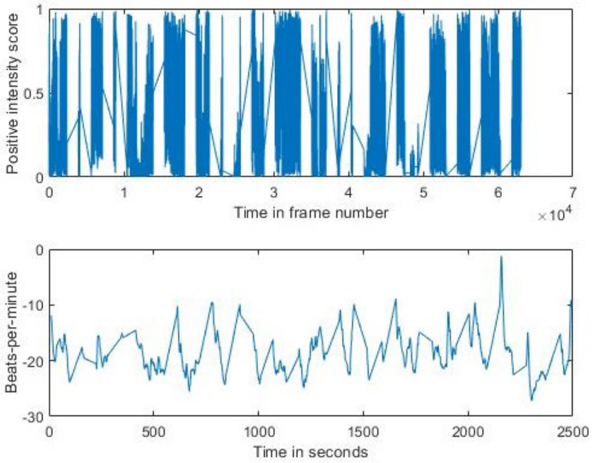


FIGURE 7. Positive intensity score of facial expression and beats-per-minute features for Subject 1.

Fig. 7 depicts positive intensity score and beats-per-minute reading plotted for a random subject's data. The features indicate that the subject's response is random throughout the viewing session and is not affected by the sequencing of the stimuli clips.

A. FACIAL EXPRESSION (FE) FEATURES

Facial expression recognition (FER) is achieved using a system that has been trained using non-laboratory based data from World Wide Web, news, and TV programs [51]. This system uses Viola-Jones algorithms to detect facial region and extract 48 landmark points. The landmark points are used to extract a number of geometric features. The face regions are used to extract a number of scale-invariant feature transform (SIFT) features which are further dimensionally reduced by the minimum redundancy maximum relevance algorithm. A feature-level fusion is conducted between the geometric features and the SIFT features, which is classified into one of the three emotion categories (i.e., positive, negative, neutral) with a three-class SVM classifier. The three emotion classes are specified from prototypical emotions (i.e., positive: happiness, surprise; negative: fear, sadness, and anger) [6].

The outputs of the FER system are frame-by-frame intensity scores for positive (POS), negative (NEG), and neutral (NEU) emotion categories. These scores indicate the probability that a particular frame would be POS, NEG, or NEU ($P^{POS,NEG,NEU}$). No further normalization is applied on these scores as they are already normalized (i.e., $P^{POS} + P^{NEG} + P^{NEU} = 1$). Linear interpolation is used for handling the missing frames. The raw intensity scores are smoothened with a low-pass moving average filtering with a 4-second time window. These smoothened scores are used directly as FE features. The three features are time stamped with frame number.

B. HEART RATE FEATURES

Heart rate features are extracted from timestamped beats-per-minute readings and RR intervals. Duplicate observations are discarded and missing observations between two adjacent observations are computed with linear interpolation. The low-resolution beat-per-minute readings collected in response to soccer stimuli are up-sampled from 1/3 Hz to 1 Hz to comply with tennis and movie data. The up-sampled beats-per-minute readings are then used to estimate RR intervals data. Both data are normalized by mean subtraction.

A total of 9 time-varying features are extracted from the BPM data, namely, beats-per-minute readings, variance, derivative, kurtosis, skewness, range, energies in high- and low-frequency bands, and ratio of the energies. A 4-second sliding window with 90% overlapping is used to compute the features. Derivative is computed with the first-order differences between the adjacent samples within the time window. The summations of the absolute differences are divided by the time difference of the time window to obtain the final derivative value. Statistical aggregation methods are used over the time samples within the window to compute variance, skewness, and kurtosis. Difference between the maximum and minimum of the samples is taken to compute range. Energy features are computed over all the time samples without using a sliding window. A band-pass-filter with a Kaiser window between [0.04, 0.15] Hz and [0.15, 0.4] Hz is used to compute low- (LF) and high-frequency (HF) energies. The band-passed samples are squared to compute the energy features. The ratio between these two energies (LF/HF) is computed by taking an element-by-element ratio of LF and HF. Skewness and kurtosis computed from beats-per-minute data are not used as final features due to high redundancy and invalid values.

A total of 7 HRV features are extracted from RR interval data, including RR intervals, their standard deviations (SDNN), successive RR interval pairs which differ by more than 50 milliseconds (NN50), root-mean-square difference of successive RR intervals (RMSSD), HF, LF, and LF/HF. The energy feature are computed in a similar manner for beats-per-minute data as described in the last paragraph, while the other features are computed using a 4-point sliding window. SDNN is computed by taking the standard deviation over the time samples. RMSSD was computed using a second order difference between the adjacent RR samples within the time window. A square root of the means (of the samples in time window) is taken as RMSSD. For computing NN50, a first order difference between adjacent time samples is taken, and then it is computed how many samples have difference values > 50.

C. MULTIMODAL DEEP LEARNING FEATURES

In addition to the time-varying features, this study investigates the usefulness of features extracted using deep learning approach, referred as 'deep feature' in this paper. The deep learning network is designed to contain two hidden layers and trained with a stack of (two) autoencoders. An autoencoder is an artificial neural network used for unsupervised learning of the data. Each autoencoder, a combination of encoder and decoder, learns a sparse representation in its respective hidden layer. A subsequent encoder thus extracts the representative version of the input data. In this work, the number of hidden layers and autoencoders are kept smaller due to smaller dimensionality of the input data. The first autoencoder takes the extracted features (FE, BPM, HRV) as input and its output is a compressed version of the input features. These compressed features are then further encoded to extract the representative features for the input features. These representative features are then fed into the second autoencoder and the subsequent encoder in a similar manner to extract a set of deep features. Fig. 8 depicts the configuration of the deep feature extraction using the two autoencoders.

The size of each hidden layer is set to be smaller than its input size and this determines the output size of the respective autoencoder. During training each autoencoder, a number of input parameters including the size of hidden layer, number of passes, factor of regularizer (for the weights of the network), and sparsity regularizer have been configured based on heuristics. The extracted FE, BPM, and HRV features are individually fed to the first autoencoder and the outputs are successively fed to the second autoencoder. The outputs of the second autoencoder are considered and utilized as the deep features.

V. VIEWER INTEREST DETECTION

The viewer interest detection is achieved using the Gaussian mixture model (GMM) technique. GMM is a model-based technique and used in estimating interest from music and music video. Performance of such model using content-based features is found higher than support vector regression [52], [53]. This study chooses GMM over traditional classifiers as preliminary test results show that GMM achieve higher performance (F1-score varies 61-64% for domain-general approach) than Adaboost, SVM, and Decision Tree (F1-score varies 32-61% for domain-general approach) in this experimental context.

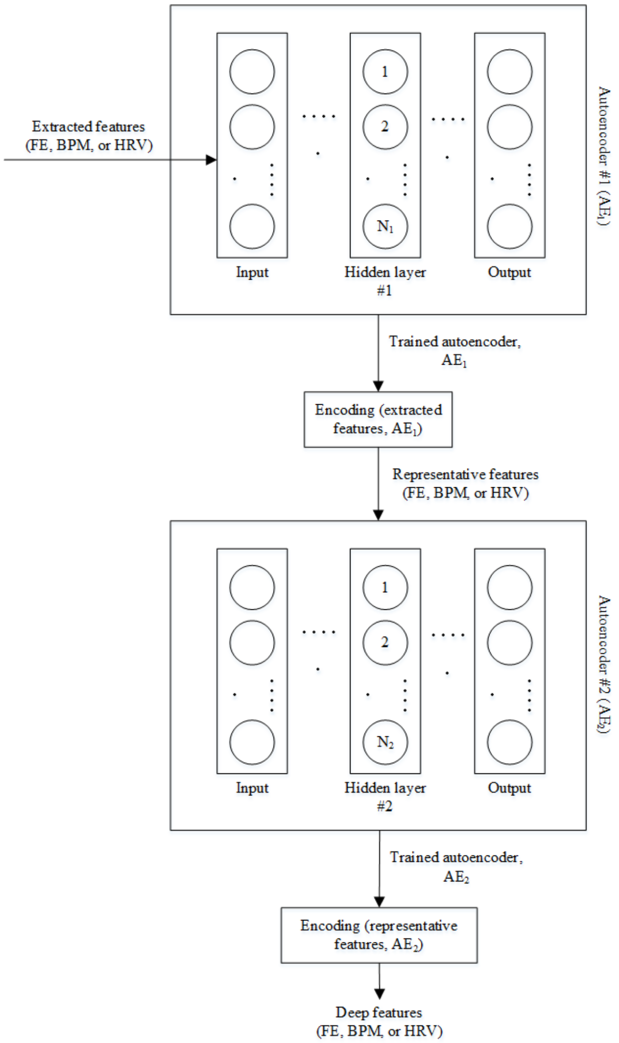


FIGURE 8. Deep feature extraction from FE, BPM, and HRV modalities using autoencoders. The sizes of the hidden layer for the first autoencoder (N_1) across the three modalities are: FE = 2; BPM = 5; and HRV = 7. Similarly, sizes of the hidden layer for the second autoencoder (N_2) are: FE = 1; BPM = 3; and HRV = 4.

The detection method uses two GMM models trained with separate feature samples labelled as 'interesting' and 'not-interesting', to classify the features. During testing, a set (of samples) is tested against the two GMMs using a posterior probability function. The posterior function in MATLAB produces a probability score (i.e., posterior probability of each GMM component) and a likelihood score (i.e., negative log-likelihood). Equations (1) and (2) show the tests where δ_{INT} and δ_{NINT} are the two GMMs and S is the set of samples being tested. The posterior probabilities and likelihood scores are denoted respectively as Φ and Ψ.

$$(\Phi_{INT}, \Psi_{INT}) = Test(\delta_{INT}, S) \quad (1)$$

$$(\Phi_{NINT}, \Psi_{NINT}) = Test(\delta_{NINT}, S) \quad (2)$$

A decision on predicted label is taken using the maximum posteriori classification approach, where the label of the GMM having a higher likelihood score is considered as the predicted label, LPREDICT (shown in (3)). The GMMs are trained and tested with each of FE, BPM, and HRV channels (of features) individually to check each channel's individual performance. A principal component analysis (PCA) is not included since our preliminary tests revealed that the principal components perform poorly if they are treated as features and applied in classification. The PCA is applied separately over the features from each channel and the first 3 principal components were taken to use as feature for classification. The accuracy achieved was varying between 40-41% across the three classification approaches.

$$L_{PREDICT} = argmax(\Psi_{INT}, \Psi_{NINT}) \tag{3}$$

The features (from three channels) are trained and tested in identical manner using three approaches, namely, domain-general, cross-domain, and domain-specific approaches (described in Table 3). A subject-dependent iteration is applied in this cross-validation to check if there is any possible subjective bias over the detection performance.

TABLE 3. Classification Approaches Applied for Viewer Interest Detection – Soccer, Tennis, and Movie Data are Denoted Respectively as D_s, D_t, and D_m.

Approach	Train vs Test Data
Domain-general	$D_S+D_T+D_M$ vs D_S, D_T, and D_M
Cross-domain	D_S vs D_M; D_M vs D_S; D_T vs D_M; and D_M vs D_T
Domain-specific	D_S vs D_S; D_T vs D_T; and D_M vs D_M

General assumption: Let us consider that T_s and T_r are two matrices of FE, BPM, or HRV features for testing and training respectively. The number of subjects for sports (i.e., 12) and movie (i.e., 20) data is denoted by t, where $t \in [12, 20]$. Assume that $(T_s\ , T_r) = [X_1, X_2, \ldots, X_t]$, where X is the normalized feature matrix for a random subject and t denotes the number of subjects. $X = (x_{ij})_{m\times n}$ where x_{ij} is the ith instance of the jth feature, and $Y = [y_1, y_2, \ldots, y_m]$: $y_j = [0, 1]$ is the subject-independent ground truth labels, where '0' indicates 'interesting' and '1' indicates 'not-interesting'.

A. DOMAIN-GENERAL APPROACH

This classification approach tests data obtained for soccer, tennis, and movie clips separately against a model trained with all of them. Here 'data' denotes the selected 'features'. In this approach, the features extracted from both the movie data are separately tested against a model trained with both movie and sports (soccer and tennis) data. A subject-independent ground truth and a leave-one-video-out cross-validation method have been used. Due to a shorter length, data from the 20 movie clips has been considered as a single data unit during classification. Thus, the leave-video-out cross-validation distributes the data from 3 soccer, 2 tennis, and 20 movie clips in such forms: vid_1 (soccer); vid_2 (soccer); vid_3 (soccer); vid_4 (tennis); vid_5 (tennis); and vid_6 (all movie clips).

Initialization: The features in T_s come from a single video data ($vid_{k\in[1,2,\ldots,6]}$), while features in T_r are from the remaining video data ($vid_{l=[1,2,\ldots,6]-k}$). The features in T_s and T_r are separately normalized so that they have zero mean and unit variance.

Step 1: The feature instances in $T_r = [X_1, \ldots, X_t]$ are separated into 'interesting' and 'not-interesting' clusters (i.e., α_{INT}, α_{NINT}) using ground-truth labels in Y. For this purpose, a direct mapping between time-based ground truth labels (y_i) and time-stamped feature instances (x_{ij}) is obtained, as $[x_{ij}, y_i]$.

Step 2: The separated training instances are used to train an 'interesting' GMM (δ_{INT}) and a 'not-interesting' GMM (δ_{NINT}), using (4) and (5). The number of components for each GMM is computed using the Akaike information criterion (AIC).

$$\delta_{INT} = \sum\nolimits_{p=1}^{c} w_p d\left(\alpha_{INT} \mid \mu_p, \sum p\right) : \alpha_{INT} \in T_r \tag{4}$$

$$\delta_{NINT} = \sum\nolimits_{p=1}^{c} w_p d\left(\alpha_{NINT} \mid \mu_p, \sum p\right) : \alpha_{NINT} \in T_r \tag{5}$$

In (4) and (5), $\mathrm{w_p}$ and d denote the weight and density of Gaussian mixture, respectively. The μ_{p}, $\sum_{\mathrm{p}}$, and C are respectively the mean, covariance, and number of components of the GMM.

Step 3: This step is iterated for each subject (t times). During testing, each subject's feature instances, $X_{t\in[12,20]} = [x_1,\ x_2, \ldots, x_{m-1},\ x_m]_n$ are further divided into segments of 'interesting' and 'not-interesting' feature instances, $[S_1, \ldots, S_r]_n$: $r < m$. Each such segment contains a number of sequential instances labelled as 'interesting' or 'not-interesting', $S = [x_1, \ldots, x_d]$: $[y_1, \ldots, y_d] \in [0 \vee 1]$.

Step 4: Each such segment, S_i is then compared against the two GMMs and a label is predicted using (1), (2), and (3). The deep features obtained for FE, BPM, and HRV channels are used in an identical manner as the original FE, BPM, and HRV features.

B. CROSS-DOMAIN APPROACH

This approach is used to confirm the need for training with new type of stimuli, which represents the key challenge in achieving a generic model. It checks whether a model trained with sports stimuli can detect viewer interest during movie stimuli, and vice versa. The training vs testing cases include: training with movie data and testing with (i) soccer and (ii) tennis data; training with (iii) soccer and (iv) tennis data and testing with movie data. This approach treats the 20 movie clips individually as separate clips rather than a whole group. A subject-dependent iterative test follows where each subject's data in response to each clip has been tested against the trained model.

Initialization: The features in T_s and T_r are obtained from the data of two different domains (sports vs movie, or movie vs sports). The features are separately normalized to have zero mean and unit variance.

Training: The features in T_r are separated into two clusters of training data labeled as 'interesting' and 'not-interesting', following the identical procedure described in Step 1 of Section V-A. Two GMMs are trained with 'interesting' and 'not-interesting' training data using (4) and (5).

Testing: For each subject, the feature data, X_i is divided into segments of feature instances labeled as 'interesting' or 'not-interesting'. The method used is identical to Step 3 of Section V-A.

Each such segment is then tested against the two GMMs and two likelihood scores are obtained, as shown in (1) and (2). The predicted label is obtained from the corresponding label of that GMM which produces a higher likelihood score as shown in (3).

C. DOMAIN-SPECIFIC APPROACH

In this approach, features extracted from the data of a specific domain (i.e., movie, soccer, or tennis) are used both in training and testing. The subject-independent ground truth and a leave-one-subject-out cross-validation is used. In case of movie data ($t = 20$), features from each subject' data are tested against the model trained with the remaining 19 subjects' features. And in case of soccer and tennis data ($t = 12$), features from each subject's data are tested against the model trained with the remaining 11 subjects' feature data.

Initialization: The leave-one-subject-out cross-validation assumes $T_s = X_{i=1:t}$ as a random subject's feature data and $T_r = [X_1, \ldots, X_t] - X_i$: $t \in [12, 20]$ as the remaining subjects' (12 or 20) feature data. The features in T_s and T_r are normalized separately to have zero mean and unit variance.

Classification: The procedure is identical to the steps described in Section V-A. The feature instances in T_r are separated and used to train two GMMs using (4) and (5). During testing, the feature instances in T_s are divided into segments of feature instances and each of them is tested against the two GMMs using (1) and (2). The predicted label is obtained from the label of the corresponding GMM that yields a higher likelihood score, using (3).

VI. EXPERIMENTAL RESULTS

A. PERFORMANCE MEASURES

Each time a predicted label ($L_{PREDICT}$) is computed, it is compared with the actual ground truth label, L_{ACTUAL}. L_{ACTUAL} is calculated as $[y_1 \vee y_2 \vee \ldots \vee y_d]$: $y_i = [0 \text{ or } 1]$. Four performance matrices including true positive, false positive, true negative, and false negative (i.e., respectively TP, FP, TN, and FN) are computed. True positive is measured as the number of feature (instance) segments which are both predicted and labelled as 'interesting'. Similarly, false positive is measured as the number of segments which are predicted as 'interesting' but labelled as 'not-interesting'. True negative is the number of segments which are both predicted and labelled as 'not-interesting'. False negative is measured as the number of segments which are predicted as 'not-interesting' but labelled as 'interesting'. The performance matrices are used further to compute recall, specification, and F1-score as performance measures.

The total counts for TP, FP, TN, and FN of each subject's feature data are combined as confusion matrix. These subject-dependent confusion matrices are further aggregated into a summarized confusion matrix for each channel (i.e., FE, BPM, and HRV), presented in Table V. The aggregation is completed by taking the cases with the maximum TP and maximum TN. The precision, recall, and F1 scores obtained during classification in a subject-dependent manner are combined for each video/clip and then combined for each channel. Then means, standard deviations, and maximums of these precision, recall, and F1 scores are taken to aggregate them for each channel and for each classification approach.

TABLE 4. **Detection performance.**

Performance Indicators	Channel					
	FE		BPM		HRV	
	Prec	Rec	Prec	Rec	Prec	Rec
Domain-general Approach						
Mean	49	57	49	**64**	**50**	39
SD	2	7	11	32	26	28
Max	50	63	61	**94**	**84**	75
Cross-domain Approach						
Mean	47	51	47	**58**	**48**	21
SD	1	4	10	20	27	11
Max	50	57	55	**74**	**81**	38
Domain-specific Approach						
Mean	51	54	50	**78**	**60**	71
SD	6	8	18	34	12	30
Max	65	69	70	83	**87**	**96**
Domain-general Approach with Deep features						
Mean	50	80	52	**83**	**54**	75
SD	2	31	1	12	6	35
Max	51	**100**	53	**100**	**64**	**100**

B. ACCURACY OF VIEWER INTEREST DETECTION

Table IV presents the performance of viewer interest detection across different channels and classification approaches, indicated by the mean, max and standard deviation (SD) of precision and recall rates. Across all channels, domain-specific approach consistently produces the highest precision (78%) and recall (60%) rates, followed by domain-general and cross-domain approaches. However, use of deep features improves the performance of domain-general approach (precision: 55%, recall: 83%). The standard deviations vary within 2-30% across the three classification approaches, which indicates a critical tradeoff between higher accuracy and lower precision. Nevertheless, the maximum values of the precision-recall performance of each experiment are not that different to the mean values, indicating that the classification performance is rather consistent at 50-87% precision, and 57-96% recall rates.

HRV features consistently generates higher precision but lower recall rates than BPM and FE features across the three approaches. This indicates that HRV features give fewer but more accurate detection of 'interesting' instances.

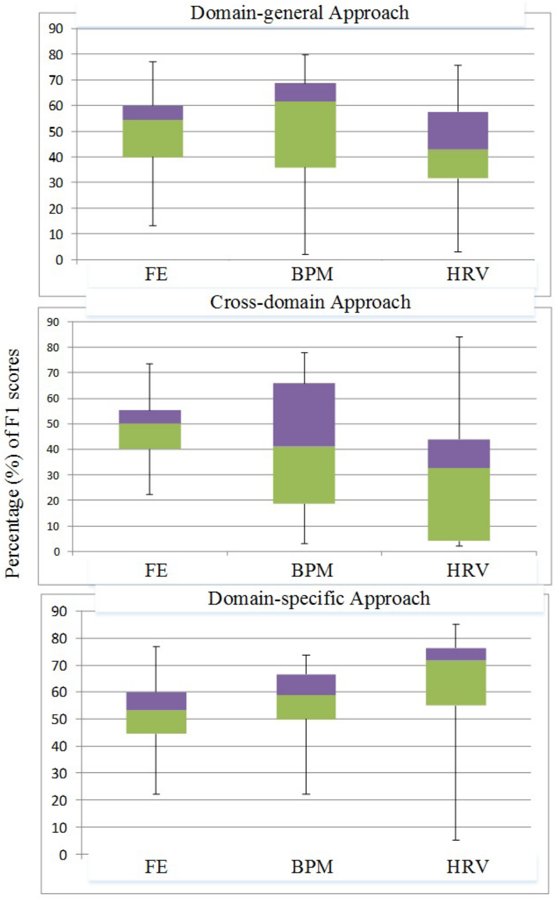


FIGURE 9. Box plot of F1-scores (in %) for facial expression, heart rate (bpm), and heart rate variability (RR) data.

However, HRV gives a large variation in precision and recall rates with standard deviations between 11-30% across the three approaches. Using BPM features generally achieves higher recall rates (more detection with less accuracy) than using HRV and FE features. The standard deviations of precisions and recall rates vary between 11-34%. The FE features have the least precision and recall rates with the lowest variance among the three feature channels.

Fig. 9 presents box plots of the F1-scores (in %) obtained for FE, BPM, and HRV features in the three classification approaches. The F1-scores obtained for all subjects in each feature channel and each classification approach are combined together to compute the minimum, 25th percentile, median, 75th percentile, and maximum. The box plots confirm the higher accuracy of domain-specific approach (45-78%) over domain-general (31-69%) and cross-domain (11-67%) approaches. It is also evident that the BPM features consistently achieve higher accuracy.

The plots show high variance as they are constructed based on a combination of each subject's data, instead of aggregation. According to the error bars, the variance in F1-scores is higher for HRV features, while lower for BPM features. The variance generally is lower for the FE features.

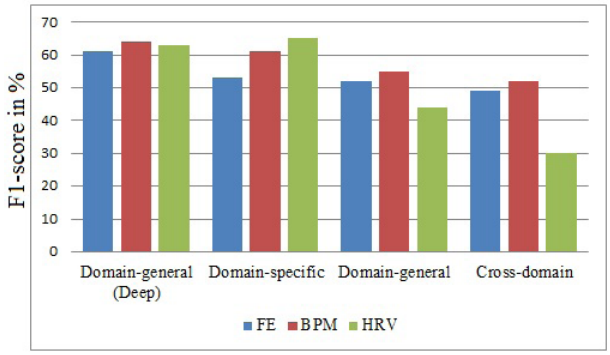


FIGURE 10. Mean F1-scores achieved by the three feature channels in domain-general (with and without deep features), domain-specific, and cross-domain approaches.

C. PERFORMANCE WITH DEEP FEATURES

Fig. 10 illustrates the means (in %) of the F1-scores obtained in the three classification approaches. The F1-scores computed for all subjects in each feature channel and each classification approach are aggregated. The use of deep features in the domain-general approach increases its performance from 45-52% to 62-64%. It achieves a higher overall accuracy than the domain-specific approach (52-65%).

Among the three feature channels, BPM features have higher and more consistently stable accuracy (51-64%) than HRV (30-65%) and FE (49-61%) features. The FE features are the second in achieving consistent accuracy. The HRV features achieve lower accuracy in the domain-general (without deep features) and cross-domain approaches, which are improved after the deep features are included. Overall, the accuracy of FE, BPM, and HRV features increases after applying deep features.

TABLE 5. Confusion matrix for best cases for each channel obtained using domain-general, cross-domain, and domain-specific approaches.

Channel	Hit	False alarm	Correct rejection
		Domain-general Approach	
FE	.5	.35	.15
BPM	**.56**	.4	.04
HRV	.5	.25	**.25**
		Cross-domain Approach	
FE	.44	.37	.13
BPM	**.5**	.36	.14
HRV	.43	.29	**.21**
		Domain-specific Approach	
FE	.5	.44	.06
BPM	**.67**	.18	.15
HRV	.5	.3	**.2**

Table V depicts the confusion matrix with cases where the TP and TN are maximum for each channel and classification approach. The hits, false alarms, and true rejections are computed by taking fractions of measured TP, FN, FP, and TN. The overall results indicate that the domain-specific approach performs better (0.67 hits) than the domain-general (0.56 hits) and cross-domain (0.50 hits) approaches. Both the domain-specific and domain-general approaches have zero miss rates and thus omitted in Table V.

Among the three feature channels, the use of BPM features appears to produce the most consistent and highest detection rates in all three classification approaches. Using HRV features consistently give higher rejection rates than FE and BPM features. The (high resolution) beats-per-minute and RR interval data appear to produce better features than facial expression features. This also indicates that the BPM features are better in detecting an 'interesting' state while the HRV features are better in detecting a 'not-interesting' state.

VII. DISCUSSION

This work uses machine learning techniques to automatically detect viewer interest in response to stimuli using facial expression and heart rate signals. Compared to facial expression, heart rate is a more a spontaneous physiological signal, therefore is expected to be more indicative to viewer interest. The experimental results support this hypothesis as heart rate (F1: 52-62%) and heart rate variability (F1: 30-64%) features show superior performance over facial expression (F1: 48-51%) features across all training-testing approaches. However, a higher variance in F1-scores when heart rate features were used could mean that facial expression is a more robust modality for detecting viewer interest, as the data is less affected by signals quality and resolution when different wearable sensors are used for recording the viewer's heart rate.

The experimental results also demonstrate that deep learning approach can improve the performance of a generic model (F1: 61-64%) up to the level of a domain-specific model (F1: 52-65%) for viewer interest detection. This supports the hypothesis that extraction of deep learning features using auto-encoder-based neural networks can capture higher-level and more representative features from viewer interest that cannot be captured by manually engineered features. However, domain-specific model still (as expected) yields the highest accuracy compared to the domain-general and cross-domain approaches.

VIII. CONCLUSION

This work investigates the feasibility of a generic model to detect viewer interest using facial expression and heart rate features. To overcome the constraint of limited modality, the experiment includes features from three channels including facial expression, heart rate, and heart rate variability. Heart rate and variability features achieve performance with a diverse range, which can be minimized by using higher resolution data. Deep learning features are useful to capture subtle and representative patterns in viewer bio-signals, which facilitates building a generic model for affect recognition. The size of the neural network is kept small in this work (i.e., two auto-encoders) since the dimensionality (i.e., size) of the extracted features is relatively low. A higher sample size would be useful in designing a bigger and sophisticated neural network. Future work will investigate more sophisticated nonlinear features including Poincare and recurrence features. Other classification techniques can be used (apart from GMM) to test if the performance of the general model can be improved further.

REFERENCES

[1] P. J. Silvia, "Interest and interests: The psychology of constructive capriciousness," *Rev. Gen. Psychol.*, vol. 5, no. 3, pp. 270–290, 2001.

[2] M. Ainley, S. Hidi, and D. Berndorff, "Interest, learning, and the psychological processes that mediate their relationship," *J. Educ. Psychol.*, vol. 94, no. 3, pp. 545–561, 2002.

[3] E. S.-H. Tan, "Film-induced affect as a witness emotion," *Poetics*, vol. 23, nos. 1–2, pp. 7–32, 1995.

[4] P. C. Ellsworth and C. A. Smith, "From appraisal to emotion: Differences among unpleasant feelings," *Motivat. Emotion*, vol. 12, no. 3, pp. 271–302, 1988.

[5] P. J. Silvia, "Interest—The curious emotion," *Current Directions Psychol. Sci.*, vol. 17, no. 1, pp. 57–60, 2008.

[6] P. R. Chakraborty, D. Tjondronegoro, L. Zhang, and V. Chandran, "Automatic identification of sports video highlights using viewer interest features," in *Proc. ACM Int. Conf. Multimedia Retr.*, New York, NY, USA, Jun. 2016, pp. 55–62.

[7] S. A. Turner, Jr., and P. J. Silvia, "Must interesting things be pleasant? A test of competing appraisal structures," *Emotion*, vol. 6, no. 4, pp. 670–674, 2006.

[8] P. J. Silvia, "Emotional responses to art: From collation and arousal to cognition and emotion," *Rev. Gen. Psychol.*, vol. 9, no. 4, pp. 342–357, 2005.

[9] P. J. Silvia, "What is interesting? Exploring the appraisal structure of interest," *Emotion*, vol. 5, no. 1, pp. 89–102, 2005.

[10] P. J. Silvia, "Cognitive appraisals and interest in visual art: Exploring an appraisal theory of aesthetic emotions," *Empirical Stud. Arts*, vol. 23, no. 2, pp. 119–133, 2005.

[11] P. Langsdorf, C. E. Izard, M. Rayias, and E. A. Hembree, "Interest expression, visual fixation, and heart rate changes in 2- and 8-month-old infants," *Develop. Psychol.*, vol. 19, no. 3, pp. 375–386, 1983.

[12] H. Grabner, F. Nater, M. Druey, and L. van Gool, "Visual interestingness in image sequences," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, Oct. 2013, pp. 1017–1026.

[13] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, Jun. 2014.

[14] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2085–2098, Dec. 2014.

[15] G. Zen, P. de Juan, Y. Song, and A. Jaimes, "Mouse activity as an indicator of interestingness in video," in *Proc. ACM Int. Conf. Multimedia Retr.*, New York, NY, USA, Jun. 2016, pp. 47–54.

[16] W.-T. Peng *et al.*, "Editing by viewing: Automatic home video summarization by viewing behavior analysis," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 539–550, Jun. 2011.

[17] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 15–28, Jan./Mar. 2017.

[18] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proc. ACM Int. Conf. Image Video Retr.*, Santorini, Greece, Jul. 2009, p. 31.

[19] W.-T. Peng *et al.*, "A user experience model for home video summarization," in *Proc. Int. Conf. Multimedia Modeling*. Heidelberg, Germany: Springer, Jan. 2009, pp. 484–495.

[20] I. Arapakis, I. Konstas, and J. M. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, Oct. 2009, pp. 461–470.

[21] S. Wang, Z. Liu, Y. Zhu, M. He, X. Chen, and Q. Ji, "Implicit video emotion tagging from audiences' facial expression," *Multimedia Tools Appl.*, vol. 74, no. 13, pp. 4679–4706, 2015.

[22] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 505–523, 2011.

[23] X. Huang, A. Dhall, R. Goecke, M. Pietikainen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2706–2721, Oct. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8323249/

[24] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 391–400, Feb. 2013.

[25] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, 1993.

[26] R. Gupta, M. K. Abadi, J. A. C. Cabré, F. Morreale, T. H. Falk, and N. Sebs, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proc. ACM Int. Conf. Multimedia Retr.*, New York, NY, USA, Jun. 2016, pp. 317–320.

[27] M. Soleymani and M. Pantic, "Multimedia implicit tagging using EEG signals," in *Proc. IEEE Int. Conf. Multimedia Expo*, San Jose, CA, USA, Jul. 2013, pp. 1–6.

[28] J. Han, X. Ji, X. Hu, L. Guo, and T. Liu, "Arousal recognition using audio-visual features and FMRI-based brain response," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 337–347, Oct. 2015.

[29] P. R. Chakraborty, L. Zhang, D. Tjondronegoro, and V. Chandran, "Using viewer's facial expression and heart rate for sports video highlights detection," in *Proc. ACM Int. Conf. Multimedia Retr.*, Shanghai, China, Jun. 2015, pp. 371–378.

[30] C. A. Smith, "Dimensions of appraisal and physiological response in emotion," *J. Pers. Social Psychol.*, vol. 56, no. 3, pp. 339–353, 1989.

[31] K.-M. Ong and W. Kameyama, "Video summary based on human bio-signals—Evaluation of its applicability to first-time viewers," in *Proc. 2nd Int. Symp. Aware Comput.*, Tainan, Taiwan, Nov. 2010, pp. 251–256.

[32] A. Lang, J. Newhagen, and B. Reeves, "Negative video as structure: Emotion, attention, capacity, and memory," *J. Broadcast. Electron. Media*, vol. 40, no. 4, pp. 460–477, 1996.

[33] A. G. Money and H. Agius, "Analysing user physiological responses for affective video summarisation," *Displays*, vol. 30, no. 2, pp. 59–70, 2009.

[34] A. G. Money and H. Agius, "ELVIS: Entertainment-led video summaries," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 6, no. 3, p. 17, 2010.

[35] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *Int. J. Semantic Comput.*, vol. 3, no. 2, pp. 235–254, 2009.

[36] J. Reeve, "The face of interest," *Motivat. Emotion*, vol. 17, no. 4, pp. 353–375, 1993.

[37] G. W. Alpers, D. Adolph, and P. Pauli, "Emotional scenes and facial expressions elicit different psychophysiological responses," *Int. J. Psychophysiol.*, vol. 80, no. 3, pp. 173–181, 2011.

[38] M. G. Bos, P. Jentgens, T. Beckers, and M. Kindt, "Psychophysiological response patterns to affective film stimuli," *PLoS ONE*, vol. 8, no. 4, p. e62661, 2013.

[39] J. Fleureau, P. Guillotel, and Q. Huynh-Thu, "Physiological-based affect event detector for entertainment video applications," *IEEE Trans. Affective Comput.*, vol. 3, no. 3, pp. 379–385, Jul. 2012.

[40] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Proc. 10th IEEE Int. Symp. Multimedia*, Berkeley, CA, USA, Dec. 2008, pp. 228–235.

[41] J. De Jonckheere, D. Rommel, J. L. Nandrino, M. Jeanne, and R. Logier, "Heart rate variability analysis as an index of emotion regulation processes: Interest of the Analgesia Nociception Index (ANI)," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, San Diego, CA, USA, Sep. 2012, pp. 3432–3435.

[42] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony?" *Biofeedback*, vol. 35, no. 2, pp. 54–61, 2007.

[43] A.-C. Lin, F.-Y. Yen, M.-H. Sun, F.-S. Shang, S.-T. Tang, and J.-H. Lin, "A real-time portable analyzer for anger emotion," in *Proc. Int. Conf. Electron. Inf. Eng.*, Kyoto, Japan, Aug. 2010, pp. 92–95.

[44] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3687–3691.

[45] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, Jul. 2008, pp. 160–167.

[46] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 675–678.

[47] P. R. Chakraborty, "Detecting viewer interest in video using facial and heart rate responses," Ph.D. dissertation, Queensland Univ. Technol., Brisbane, QLD, Australia, 2017.

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

[49] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 43–55, Jan. 2015.

[50] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

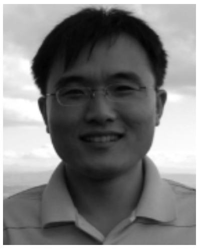
[51] L. Zhang, D. Tjondronegoro, and V. Chandran, "Facial expression recognition experiments with data from television broadcasts and the World Wide Web," *Image Vis. Comput.*, vol. 32, no. 2, pp. 107–119, 2014.

[52] J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, and H.-M. Wang, "The acousticvisual emotion Guassians model for automatic generation of music video," in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, Oct. 2012, pp. 1379–1380.

[53] J. C. Wang, Y. H. Yang, H. M. Wang, and S. K. Jeng, "Modeling the affective content of music with a Gaussian mixture model," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 56–68, Jan. 2015.

PRITHWI RAJ CHAKRABORTY received the bachelor's degree in computer science and engineering from the Chittagong University of Engineering and Technology, Bangladesh, and the Ph.D. degree in information systems from the Queensland University of Technology, Brisbane, Australia. He was a Research Fellow with Southern Cross University, where he is currently an Academician. His primary research interests include affective computing, human–computer interaction, affective multimedia, and emotion analysis.

DIAN WIRAWAN TJONDRONEGORO (M'12–SM'15) is currently a Professor and the IT Discipline Leader with the School of Business and Tourism, Southern Cross University. He has been innovating mobile health/wellness apps as part of his research with the Young and Well CRC, NHMRC partnership, Diabetes Australia and ARC Discovery. He has published over 130 papers in top-tier peer-reviewed journals and conference proceedings, including the IEEE TAC, *Neurocomputing*, IEEE T-MM, ACM TOMM, and ACM MM. His passion is in applied research and real-world teaching to make impacts in ubiquitous computing and multimedia fields. His current research interests include personalized health technology using multimedia big data in IoT framework. His expertise is in affective computing (facial expression recognition and physiological signals analysis), video analysis and communication, and mobile health applications. He is an Associate Editor of the IEEE Access and a Regular Reviewer for prestigious journals, such as the IEEE T-MM, IEEE TCSVT, ACM MTA, and PR.



LIGANG ZHANG received the Ph.D. degree from the Queensland University of Technology in 2012. He is currently a Senior Research Officer with the School of Engineering and Technology, Central Queensland University, Australia. He has published a book in Springer and over 30 peer-viewed papers in top tier journals and conference proceedings. His research interests include facial expression recognition, affective computing, image segmentation and recognition, and computer vision. He serves in the Technical Program Committees for over 20 international conferences.

VINOD CHANDRAN (S'85–M'90–SM'01) received the bachelor's degree in electrical engineering from IIT Madras, the M.S. degree in electrical engineering from Texas Tech University, the M.S. degree in computer science from Washington State University, and the Ph.D. degree in electrical and computer engineering from Washington State University in 1990. He is currently an Adjunct Professor with the Queensland University of Technology, Australia. He has supervised 14 Ph.D. students as the Principal Supervisor and 20 post-graduate research students to completion as an Associate Supervisor. He has authored or co-authored over 180 journal and conference papers. His research contributions span signal processing, image processing, and pattern recognition with applications mainly to biometrics and biomedical systems. He is a member of the Australian Computer Society.