Received January 3, 2019, accepted March 4, 2019, date of publication April 15, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911235

EmoWare: A Context-Aware Framework for Personalized Video Recommendation Using Affective Video Sequences

ABHISHEK TRIPATHI[1], (Student Member, IEEE), T. S. ASHWIN[2], (Student Member, IEEE), AND RAM MOHANA REDDY GUDDETI[2], (Senior Member, IEEE)

[1]Rivigo Services Pvt. Ltd., Bengaluru, India
[2]Department of Information Technology, National Institute of Technology Karnataka, Mangalore 575025, India

Corresponding author: Abhishek Tripathi (abhishek.tripathi2421@gmail.com)

ABSTRACT With the exponential growth in areas of machine intelligence, the world has witnessed promising solutions to the personalized content recommendation. The ability of interactive learning agents to make optimal decisions in dynamic environments has been proven and very well conceptualized by reinforcement learning (RL). The learning characteristics of deep-bidirectional recurrent neural networks (DBRNN) in both positive and negative time directions has shown exceptional performance as generative models to generate sequential data in supervised learning tasks. In this paper, we harness the potential of the said two techniques and propose EmoWare (emotion-aware), a personalized, emotionally intelligent video recommendation engine, employing a novel context-aware collaborative filtering approach, where the intensity of users' spontaneous non-verbal emotional response toward the recommended video is captured through interactions and facial expressions analysis for decision-making and video corpus evolution with real-time feedback streams. To account for users' multidimensional nature in the formulation of optimal policies, RL-scenarios are enrolled using on-policy (SARSA) and off-policy (Q-learning) temporal-difference learning techniques, which are used to train DBRNN to learn contextual patterns and to generate new video sequences for the recommendation. System evaluation for a month with real users shows that the EmoWare outperforms the state-of-the-art methods and models users' emotional preferences very well with stable convergence.

INDEX TERMS Reinforcement learning, Q-learning, SARSA, deep bidirectoinal recurrent neural network, multi-armed bandit, video recommendation, affectiva, intensities of emotions, emotion-based information retrieval.

I. INTRODUCTION

Affective computing is an academic discipline which is empowering machines to acquire human-like characteristics. Among many non-verbal and involuntary channels through which humans express themselves, facial expressions hold paramount importance. Studies such as [31], [32] have shown that Human-Machine Affective Interaction has been proving considerable results in learning and responding towards affective stimuli in natural and uncontrolled environments. If such learning agents are trained congruously, the performance of the recommendation systems can greatly be improved.

Automatic video recommendation is one of those considerable challenges. A lot of video sharing platforms with admirable recommendation algorithms such as YouTube's deep neural nets [9], Netflix's recommender system [10], etc., have been devised to deliver preferred content to the user but the growing number of applications are not just limited to serve the purpose of entertainment. Recommender systems are constantly getting used in domains such as eLearning, Healthcare, etc., for education, streamlining communications, monitoring psychological behavior of patients in music & video based therapies, in psychotherapies, in multi-modal systems etc., where recommendations heavily rely on users' history and other metadata. However, limitations still exist if the decision making is considered in real time.

It has been observed that search criteria of users are largely moods and emotional state driven [33]–[35], hence their preferences change over time with a high degree of fluctuations. For instance, consider a viewer who is in a happy mood and searching for happy/joyful videos. Unless the viewer

The associate editor coordinating the review of this manuscript and approving it for publication was Omid Kavehei.

explicitly knows what to search, according to the intensity of the happy emotional state, a system might not suggest apt content to reconcile the viewer's expectations. Similarly, imagine a viewer exploring scary movies and without knowing the extent of horror among his/her shortlisted content, he/she ends up watching one which was not scary for him/her at all. We also strongly believe that preferences have a logical and inevitable dependency on context with possibilities to have short-term trends but a recurring seasonality. For example, viewers can have totally different behavior during weekends as compared to working days, routine work can trigger repeated sequential patterns in preferences, different parts of a day can have different moods and hence different preferences, etc. *So the problem can be condensed as*: ''Can personalized recommendation get smart enough to identify and serve the exact content''?

There has been a lot of advancement in emotion-based information and data retrieval such as music [36], [37], images [38] where agents make decisions to suit the user's mood efficiently. However, to the best of our knowledge, previous studies have mainly focused on training models or taking decisions based on global emotion label of previously visited contents, either considering only a few previous interactions or using global watched history to identify the domain of interest. But no automatic video recommendation systems take into account the intensity of different emotions present in videos, the emotional impact created on user, taking decisions based on streaming facial expressions data (throughout video length), the temporal patterns in which the videos are searched and watched along with changing contextual information of users in real time.

Asserting that recommendation is a continuous optimization process, in this work, we incorporate the temporal human nature, highlight the impact of intensities of different emotions and give a comparative and in-depth study of the learning trends of two coupled (RNN powered RL) algorithms namely RNN+SARSA and RNN+Q-Learning to solve the above-stated problem. They are designated so because the optimal policies generated by RL algorithms (SARSA and Q-Learning) are used by RNN for state sequence learning and generating new sequences which are again used by RL for optimization. The framework is an instance of a context-aware collaborative filtering approach to solve the present video recommendation systems' shortcomings using real-time algorithms. Our system slowly evolves its Dataset using feedback collected from multiple users into more precise annotations to learn general and specific users' behavior w.r.t emotions. Evaluation of the system based on the proposed methodology demonstrates that RNN+Q-Learning algorithm is effective in learning exploratory-users behavior while RNN+SARSA algorithm performs well in modeling exploitative-users.

To summarize, the major contributions of this paper which differ from existing work, are as follows:

- Firstly, unlike the decoupled behavior of RNN and RL, the hybrid employed algorithm can best capture the long-term context-aware sequential information to model states representation for RL which in turn provides best possible short-term dynamic behavior's sequence representation to RNN.
- Secondly, the video annotation technique provides continuous affective annotations on basic emotions from a large number of users through crowd-sourcing.
- Thirdly, the implicit feedback mechanism through live video streams not only provides annotations for the system's initial dataset but also helps to identify video segments with strong emotions. Thus strengthening the system for video summarization.
- Finally, the introduction of a dynamic mood based and context-aware recommendation system with a new criterion for searching and filtering based on emotional intensity in videos.

Organization of the rest of the paper is as follows. Section II provides the related work on existing affective computational models. Section III highlights the main system constituents and describes the proposed EmoWare's framework in detail. Section IV explicates the experimental setup, obtained results, and analysis. Finally, in Section V, we conclude the paper with future directions.

II. RELATED WORK

Research on dynamic user modeling for affective recommendation exists for various applications. Ardissono and Torasso [49] proposed an unobtrusive user behavior analysis system for dynamic user modeling in web store shells. In [50], ontology-based multi-agent dynamic user profile modeling is performed using short-term and long-term interests. Further, social media systems, considered by Yin *et al.* [48] for dynamic user modeling and recommendation, used intrinsic interests and temporal context-aware mixture models. But these work were neither designed for multimedia content recommendation systems nor used affective features for the recommendation.

Li *et al.* [51] considered the e-commerce scenario and proposed a hybrid encoder using a neural networks framework for session-based recommendation system. But the number of item attributes considered were very less with the absence of some of the very important ones including price relations and item categories. Quadrana *et al.* [57] introduced a hierarchical recurrent neural networks architecture for personalizing session-based recommendations using cross-session information transfer but the absence of automatic feature extraction for model training was a major drawback.

Moreover, there are several studies on sequential data modeling too. Smironova and Vasile [16] proposed sequential modeling for recommendation using contextual recurrent neural networks but the approach cannot be extended to real-time video recommendation since they performed the test on the YouChoose dataset. Donkers *et al.* [52] used gated recurrent unit to optimize the sequential user-based recommendations but failed to deliver acceptable performance if the user's consumption sequence is short. Li *et al.* [17] proposed

TABLE 1. Summary of existing works.

Authors	Methods	Merits	Limitations	Multimedia Content Used	Affective Representation Used
Reynolds et al. [42]	Personalized automatic music playlist generation	Features like location, activity, temperature are used	Only contextual and environmental information is used	Audio	No
Hu et al. [43]	NEXTONE Player, a music recommender system	Forgetting curve and favouredness are used for music recommendation	Only user behavior aspects are considered.	Audio	No
King et al. [44]	Music playlist generation using hierarchical clustering and Q-Learning	Reinforcement learning over hierarchically-clustered sets of songs is used to learn listening preferences.	User emotions are not considered for music recommendation	Audio	No
Choi et al. [12]	RNN based playlist generation	RNNs are trained on within track transitions	Affective content is not considered for playlist generation	Audio	No
Cardoso et al. [47]	Mood-based playlist generation	Thayer's model of mood is used as the classifier	Categorical model of moods are not considered	Audio	Dimensional
Chi et al. [2]	Emotion based automatic playlist generation	Reinforcement learning is used for recommendation	Generating playlist based on a discrete emotion label of previous 2 songs	Audio	Dimensional
Chaiarandini et al. [46]	Dynamic playlist generation	Multimedia control signals and descriptors are used	Only discrete mood bi-dimensional plane is used	Audio	Bi-dimensional
Mariappan et al. [45]	FaceFetch: emotion based multimedia content recommendation system	ProASM feature extraction techniques used and Ekman's standard emotions are used for classification	The playlist generation does not use any learning techniques to improve the playlist generation process	Audio and Video	Categorical emotions
Y et al. [5]	XV-Pod and affective video player	Bodymedia sensewear is used to analyse user's emotion	Only physiological signals are used for emotion classification	Video	Physiological signals
Covington et al. [9]	Youtube video recommendation system	Deep neural networks are used for candidate generation and video ranking	Affective content analysis is not done for the recommendation	Video	No
Carlos et al. [10]	Netflix recommender algorithm	Features such as Trending content, video similarity, evidence selection are used	Affective content analysis is not done for the recommendation	Video	No
Zhao et al. [64]	Facial expression analysis for video classification and recommendation	Used Haar-like features for spatial and HCRFs for temporal feature fusion	Reinforcement aspect is not considered	Video	Categorical emotions

a hybrid model that combines deep q-network with LSTM to learn the representation of hidden states. Berglund *et al.* [14] proposed probabilistic interpretations of bidirectional RNNs that is used to reconstruct missing gaps efficiently with complex dynamics. Ko *et al.* [53] used both collaborative filtering and language modeling for applications like music and mobility prediction using collaborative recurrent neural networks. But these work did not perform emotion/mood analysis for dynamic user recommendation.

Considering neuro-dynamic environments, Simkins *et al.* [4] used reinforcement learning technique to derive behavior from personality where they used trait-theoretic personality models as reinforcement learning agents. But they tested only on Atkinson's computational models. Olabiyi *et al.* [15] used bidirectional RNNs for driver action prediction. Zhao *et al.* [54] proposed deep recommender system framework using pair-wise deep reinforcement learning for recommendations with negative feedback. Temporal orders and positive feedback are not at all considered for recommendation. Wang *et al.* [55] proposed exploration and exploitation trade-off as a reinforcement learning task for personalized music recommendation. But most of the above-mentioned works did not consider the affective components in the design of their recommendation system.

In the comprehensive domain of multimedia content recommendation, several works [3], [55], [56] introducing various state-of-the-art techniques have been published where this broad domain is targeted with varied multimedia contents and affective state representations, we summarize some of the key existing work along with their merits and limitations on dynamic user modeling for multimedia content recommendation system in Table 1. Other methods include, a change in emotional intensity before and after watching videos [5], building user personality models [4], video affective content analysis including spontaneous response and emotional descriptors [3] etc. Furthermore, the effect of induced emotions in viewers plays a vital role in affective modeling and plenty of research on automatic machine generated predictions for ''intended elicited emotions in humans''

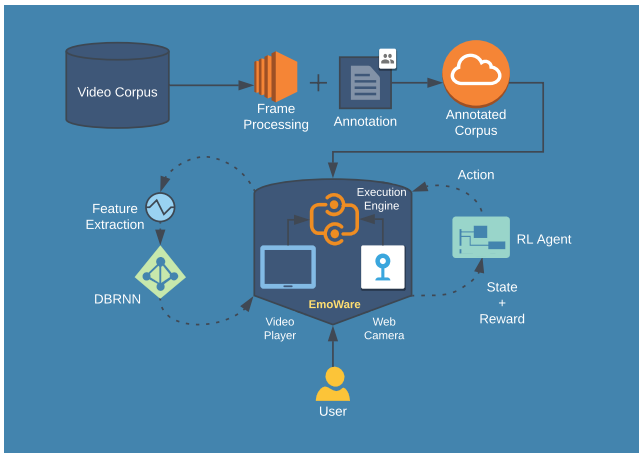

FIGURE 1. EmoWare's proposed system architecture.

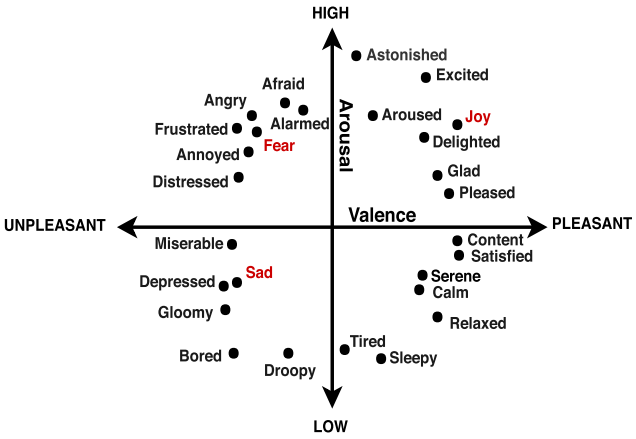

FIGURE 2. Target emotions in 2D valence-arousal circumplex.

has although accomplished many milestones using audio-visual features [25], [26], implicit video emotion tagging using audiences' facial expression though Bayesian Network [65], affective image analysis based on perception subjectivity [67]–[70], scientific analysis of speech in phonetics such as in Praat [28], text-based extractions of tones and emotions such as IBM Watson's Tone Analyzer [27], but these techniques have certain drawbacks relative to approaches and techniques such as long-term context-aware information extraction, effect of continuous affective dynamics, reinforcing factors using implicit feedback mechanisms and real-time decision making using contextual information.

So to address the aforementioned limitations, we proposed EmoWare, a dynamic user modeling framework for video content recommendation using recurrent neural networks and reinforcement learning.

III. EMO WARE

EmoWare (Emotion-Aware), is an emotional intelligence driven video recommendation system, by far, the first of its kind. In this section, we first accord its system overview followed by enclosing its target emotion domain. Next, we formally construct the target problem into a generic framework and finally, unfold the solution through this novel hybrid optimization model with the description of all the constituent elements.

A. SYSTEM OVERVIEW

The overall architecture and flow of the proposed recommendation system are illustrated in Fig. 1. EmoWare is composed of two main constituents: one for learning and adapting short-term behavior and another for learning long-term trends.

Reinforcement Learning agent (RL-Agent) captures inputs from the user's activity using data provided by video-player and camera feed to formulate policies in coherence with the behavior shown by the user and to retrieve a small subset of video from a large video corpus. These videos are generally relevant to the dynamic real-time behavior shown by the user to the system. Here, a broad level of personalization is achieved via collaborative filtering.

In order to make the system more context-aware to provide an appropriate personalized recommendation, relative importance is learned based on patterns and trends from user's feedback history. The task is accomplished by providing a rich set of contextual features to a deep neural network describing a user's emotional states and temporal preferences.

Execution Engine is the premier node for connecting the entire work-flow of the system. It controls all the data flow channels of feeder unit, RL-Agent, RNN as well as fetching data from video corpus and updating video annotations as per general users' feedback. The design allows live data streams to be recorded and processed in a stable and parallel fashion.

B. CONTRIVING EMOTION DOMAIN

Most of the existing works, as highlighted in section II, either represent the affective analysis through valence and arousal values or in the form of aggregate dominant emotions in the entire sample. Although, research in the affective domain using above mentioned technique reduces the overhead of identifying the core emotions values, however, in the real world, video excerpts usually contains a blend of emotions with the varying dominance of different emotions in different segments and a framework with continuous annotations using core emotion values on entire video length can be more effective in affective representation. Eminent theories and frameworks on discrete emotions such as Paul Ekman's [19] suggested that there are 6 "basic emotions": happiness, surprise, fear, sadness, anger, and disgust. Therefore, while modeling emotion domain of EmoWare, we extract out "basic" emotions from the pool of valence-arousal (V-A) circumplex [18] and consequently, joy, fear, and sadness are chosen as primary emotions from first 3 quadrants of V-A emotion plane as shown in Fig. 2 to obtain continuous annotations for each video of the corpus. Eventually, a video is represented by an emotion vector (E-Vector) embracing 3 dimensions

TABLE 2. Constituents of each component of E-Vector.

	JOY	SADNESS	FEAR
Emotions	joy, surprise engagement	sadness engagement	fear, surprise engagement
Expressions	smile cheekRaise lipStretch mouthOpen upperLipRaise smirk	innerBrowRaise browFurrow lipCornerDepressor noseWrinkle lipPress lipSuck eyeClosure	eyeWiden eyeClosure lidTighten lipStretch mouthOpen browRaise ignorance= (100-attention)

(Joy, Sadness, and Fear), Eqn. 1, where each dimension represents the average intensity of that emotion in the entire video.

$$\hat{E}_n = [\hat{E}_n^{(J)}, \hat{E}_n^{(S)}, \hat{E}_n^{(F)}] \tag{1}$$

where, E, S, F represent Joy, Sadness, Fear respectively and n represent sample index $1 \leq n \leq N$.

The aggregate value of each dimension presented here is guided by Affectiva [6], an emotion measurement technology company that evolved out of MIT's Media Lab, which analyzes emotions based on 33 facial points. To associate facial expressions to emotions, EMFACS mappings developed by [8] have been referred. Table 2 shows the final distribution of the features considered for each emotional component.

C. FRAMING SCENARIO INTO MARKOV MODEL

Traditional memory-based and model-based recommender systems [39] do not consider the sequence in which videos are watched by the viewers which play an important role in developing a close personalized understanding of the users, thus, providing a useful contribution in increasing the accuracy of the recommender system. To account for this factor in EmoWare, we model the sequence of videos watched by a user, as a Markov Decision Process (MDP).

An MDP is a mathematical architecture for sequential stochastic decision problem [40] which is represented by a tuple consisting of a set of finite states, a set of finite actions, a reward function, a probability or transition function, and a discount factor. For each discrete time step $t \in \{0, 1, 2, \ldots\}$, the agent always stays in a some state s_t, and among the possible actions available in state s, $a_t \in A$, the agent moves to the next state s_{t+1} by state transition function $T(s_t, a_t)$, simultaneously, receiving a reward r_t from the environment as per reward function $R(s_t, a_t, s_{t+1})$. In a certain state s, the action a is selected based on a policy $\pi(s)$. Reinforcement Learning aims to solve MDP by finding this optimal policy π that maximizes the expected cumulative reward G, known as return, through state-action value function $Q_\pi(s, a)$. The expected value of the return G obtained from episodes starting from a certain state s with the action a. $Q_\pi(s, a)$ can be

TABLE 3. Description of reward component I & II.

	Positive		Negative	
Feedback	Action	Weight	Action	Weight
Implicit	Replay Backwarding Listening Time (Full Length)	20% 10% 10%	Skipping Forwarding	-20% -10%
Explicit	Like Rating(4,5)	50% 10%,20%	Dislike Rating(1,2)	-50% -20%,-10%

expressed as follows:

$$\begin{aligned} Q_\pi(s, a) &= E_\pi\{G_t | s_t = s, a_t = a\} \\ &= E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a\right\} \end{aligned} \tag{2}$$

where, γ is the discount factor $0 < \gamma \leq 1$.

1) CONSTRUCTING STATES AND DEFINING ACTIONS

As described in Section III-B, since each video has a unique emotion vector (E-Vector) associated with it, each video is considered to be delineating a unique emotional state and the user's emotional state is assumed to be exactly resembling the state of the video being watched, independent of the previously visited videos i.e. there is a one-to-one mapping between user's and video's emotional states respectively (the degree of association of which is described in Section III-C.2), and actions denote the operations which cause transitions from one state (video) to another. Table 3 lists the set of possible actions at each state.

2) DESCRIBING REWARD FUNCTION

To evaluate the feedback from the environment, RL agents use reward function in decision making to find optimal policies. In the present scenario, the world/environment is first created by the users through watching videos (visiting states) and taking actions according to their standpoints to create a connected weighted graph which acts as input to the algorithms. To solve such MDP, EmoWare has three major components of reward function which are composed of a total of 10 actions as shown in Table 3. User feedback captured by-

- Feedback Monitor I (video player) are classified into:
  1) Implicit Feedback (constitutes $R_{1(t)}$)
  2) Explicit Feedback (constitutes $R_{2(t)}$)
- Feedback Monitor II (web cam) (constitutes $R_{3(t)}$)

The final cumulative reward function is shown by Eqn. 3:

$$R'_t = \alpha R_{1(t)} + \beta R_{2(t)} + (1 - \alpha - \beta) R_{3(t)}, \quad 0 < \alpha, \beta < 1 \tag{3}$$

where, $R_{1(t)}$, $R_{2(t)}$ and $R_{3(t)}$ are the three components of the total reward (R_t) at time t respectively. Each one of them represents the cumulative reward which is equal to the weighted

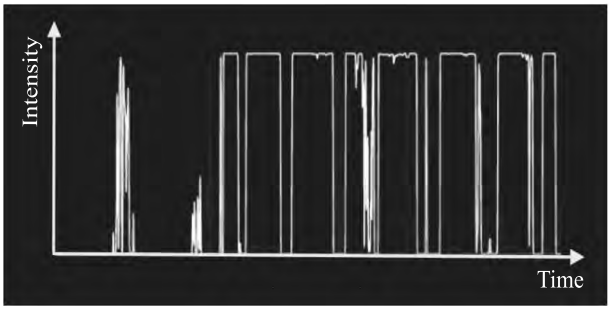

FIGURE 3. Unfiltered affective feedback of a sample video.

summation of all the action in the respective sets.

$$R_{x(t)} = \sum_{i} wa_i, \quad \forall x \in \{1, 2, 3\} \tag{4}$$

where, wa_i is the weight of the i^{th} action chosen.

Table 3 gives a detailed description of weights of each action in reward components $R_{1(t)}$ and $R_{2(t)}$ respectively. These values have been chosen empirically such that we get correct user feedback and the effects of mistakenly taken actions get nullified.

To calculate the third component of the reward function, ($R_{3(t)}$), user's emotional trends, as recorded by the system, are compared with the general trend (average trend computed from the feedback of previous viewers) of that video. But the recorded feedback is quite sensitive and can have unconventional variations w.r.t lag, level of understanding, present mood, lack of expressions, etc. as shown in Fig. 3.

Therefore, a one-to-one comparison is not possible to evaluate the obtained series with the general one. So to best evaluate the feedback, the obtained series is first filtered to smooth out short-term fluctuations and emphasize long-term trends. Since more sensitivity is needed w.r.t the most recent trends over the short term, we have chosen exponential moving average (EMA) or exponentially weighted moving average (EWMA) as the smoothing function, an infinite impulse response filter, which for a given series Y, is defined as follows:

$$S_t = \begin{cases} Y_1, & t = 1 \\ \alpha \cdot Y_t + (1 - \alpha) \cdot S_{t-1}, & t > 1 \end{cases} \tag{5}$$

where, Y_t is the value at a time period t, S_t is the value of the EMA at any time period t and coefficient α, defined as $\alpha = 2/(N + 1)$, is a constant smoothing factor between 0 and 1 representing the degree of weighting decrease. In this study, best results are produced when window size (N) is chosen to be 5, representing 5 data points of 5 seconds respectively.

Multiple segments of the video's general emotional trend are then compared with corresponding segments of the current feedback (we elucidate a segment by a pair of timestamps with non-zero emotional intensity values between them and zero outside). For comparison of two segments, we used Two-sample Kolmogorov-Smirnov (K-S) test statistic [20] to measure the goodness of fit of the feedback with the general

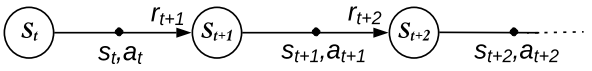

FIGURE 4. Sequence of states and actions.

one. K-S test is a non-parametric test to measure the closeness between two empirical cumulative distribution functions (ECDF), the criterion of which is defined as the maximum value of the absolute difference between two empirical cumulative distribution functions (Eqn. 6):

$$D_{m,n} = \max_{x} |S_{1,m}(x) - S_{2,n}(x)| \tag{6}$$

where, $S_{1,m}$ and $S_{2,n}$ are the empirical cumulative distribution functions of the first and second sample respectively.

The final reward value is based on the z-score of cumulative dissimilarity $D^{agg}_{(m,n)}$ value obtained from dissimilarities of all the individual segments $D^{i}_{(m,n)} \forall\, i \in [0, n]$. So lesser the z-value, more will be the positive reward and vice-versa. The standard score of a raw score x is calculated as follows:

$$z = \frac{x - \mu}{\sigma} \tag{7}$$

where, μ and σ are the mean and standard deviation of the population respectively.

D. TEMPORAL DIFFERENCE LEARNING TECHNIQUES

Temporal difference (TD) learning [41] refers to a class of model-free reinforcement learning methods which learns by bootstrapping from the current estimate of the value function. It approximates their current estimate based on some policy created through self-experience, thus an absence of environment's prior knowledge has zero impact on its performance which makes TD techniques quite suitable for learning in dynamic environments. In this problem, the world is composed of states but actions specific to each state are not predefined. They are user-formulated (once a user starts interacting with the system and visit states). So to learn a new world like this, EmoWare uses two such techniques: SARSA (an On-Policy TD Control strategy) and Q-Learning (an Off-Policy TD Control strategy).

For a random walk with an alternating sequence of states and state-action pairs, as shown in Fig. 4, in simplest form, the two algorithms, SARSA and Q-Learning are described by Eqn. 8 and Eqn. 9 and details are given by Algorithms 1 and 2 respectively.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{8}$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{9}$$

where, (s_t, a_t) and r_t represent the (state, action) pair and reward at time t respectively. r_{t+1} is the reward obtained after performing action a_t in state s_t, s_{t+1} represents the next state, α is the learning rate and γ is the discount factor. In our experiment, values of α and γ are 0.1 and 0.9 respectively.

Algorithm 1: SARSA

Initialize the Q-matrix, Q(s,a) arbitrarily;
/* for_each episode or session */
while *episodes/sessions not over* **do**
  Initialize state s;
  Choose action (a_t) from state (s_t) using ($\epsilon-greedy$) policy;
  /* for_each transition in episode */
  while *terminal state* s_t *not achieved* **do**
    Recommend next video using f(a_t);
    Observe reward r_t and new state s_{t+1};
    Choose action (a_{t+1}) from state (s_{t+1}) using ($\epsilon-greedy$) policy;
    Update Q(s_t,a_t) from (8);
    $s_t \leftarrow s_{t+1}$; $a_t \leftarrow a_{t+1}$;
  end
end

Algorithm 2: Q-Learning

Initialize the Q-matrix, Q(s,a) arbitrarily;
/* for_each episode or session */
while *episodes/sessions not over* **do**
  Initialize state s;
  /* for_each transition in episode */
  while *terminal state* s_t *not achieved* **do**
    Choose action (a_t) from state (s_t) using ($\epsilon-greedy$) policy;
    Recommend next video using f(a_t);
    Observe reward r_t and new state s_{t+1};
    Update Q(s_t,a_t) from (9);
    $s_t \leftarrow s_{t+1}$;
  end
end

1) APPROACH TO TARGET BANDIT PROBLEM

Our problem very closely resembles a multi-armed bandit problem where the objective is to find a solution which can give us maximum long-term profits by balancing exploration and exploitation. Among many approximate solutions we have used one of the semi-uniform strategies, the ϵ-greedy policy with exponential decay:

$$\epsilon_t = \epsilon_o e^{-\lambda t}, \lambda > 0 \tag{10}$$

where, ϵ_t is the ϵ value at time t and ϵ_0 is the initial value (set to 0.1 i.e. 10% exploration).

Selecting a higher ϵ at the beginning and reducing over time helps in finding the optimal action earlier and also get good long-term rewards. The decay constant λ also affects the overall average reward. In [2], the effect of different values of λ on both the algorithms are shown and 0.1 is adopted as the ideal value as it gives the maximum average reward, hence, we take the same value. Talking about function $f(a_t)$, we assume that long and sudden jumps towards high intensity of any emotion are not ideal, therefore, we use cosine similarity measure to recommend the next video whose E-Vector is among 3 nearest E-Vectors relative to present video to avoid abrupt jumps and ensure smooth transitions.

E. DBRNN TO MODEL LONG-TERM CONTEXT-AWARE BEHAVIOR

1) CONTEXTUAL FEATURES

The process of extracting out contextual features for context-aware recommendation has primarily been relying on auxiliary information sources such as user metadata (e.g. gender, age, topics of interest etc.), video metadata (e.g. genre) or session based information (e.g. click rate, location, device etc.) which have been quite promising for context representation. Having more information can be advantageous but it also increases the dimensionality of the input feature vector and can make it highly sparse. Henceforth, it's necessary to represent the knowledge that just suffices. Consequently, in our study, we consider the following features to train our system:

- Timestamp: fragmented into discrete features - hour, weekday and month.
- Emotion Type: a component of emotion from E-Vector
- Magnitude: level of intensity (High, Moderate, Low)
- Video Type: movie clip, music video

These discrete features are passed through a hash function to obtain one-hot encoded vector representations to be fed to EmoWare's learning network.

2) LONG SHORT-TERM MEMORY NETWORKS

Recurrent Neural Network is a derivative of the artificial neural network which provides a generalization of feedforward neural networks to sequences with the ability to process arbitrary sequence of inputs. RNNs have recently been observed to achieve state-of-the-art results from learning temporal sequences, be it time-series modeling, speech-recognition, etc., for generating sequential data such as in machine translation. The success of LSTM networks over conventional RNN architecture by overcoming vanishing gradient problem and learning long-term dependencies with ease unfenced another dimension of learning in many more applications. References [14] and [15] have shown that only unidirectional RNNs restricts learning ability of models by making use of only previous context.

Bidirectional RNNs can compute backward hidden sequence along with computing forward sequence in two separate hidden layers which can enrich the understanding of generative models in both the input directions. A deeper architecture posses higher learning potential, here it is achieved by stacking multiple bi-directional hidden layers on top of each other and ensuring that every hidden layer collects inputs from both forward as well as backward layers at the level just below it. But during the testing phase, we found that very deep network configuration only over-fits the data and do not contribute to the performance

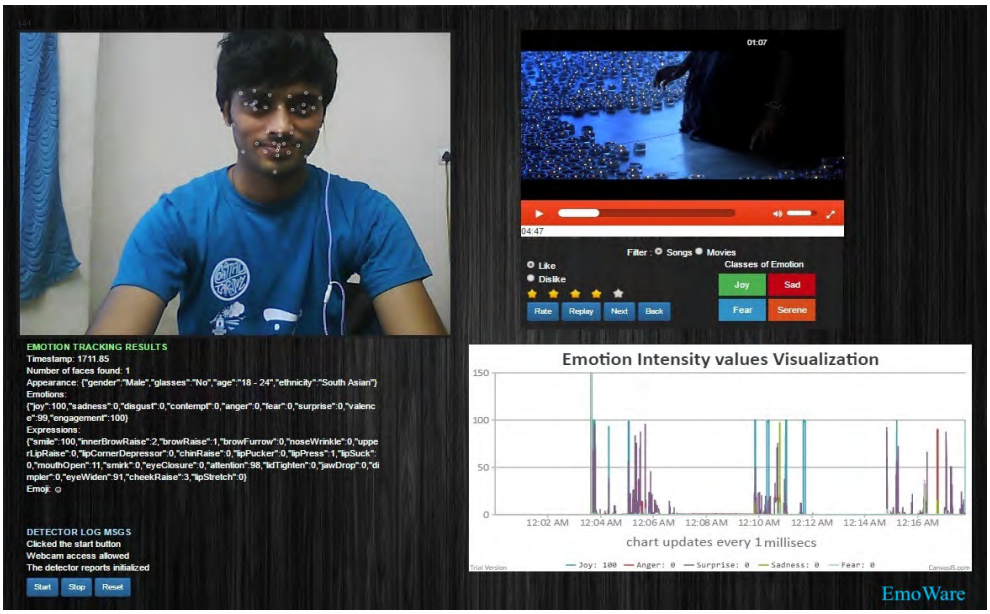

FIGURE 5. Screenshot of EmoWare's web-app interface displayed to participants.

improvement. Therefore, EmoWare's final configuration employed a 2-layer multi-cell bidirectional LSTM network.

The LSTM cell is deployed with the standard implementation [21] with changes represented by Equations 11, 12 and 13. For a given input sequence, $x = (x_1, \ldots\ldots, x_T)$, RNN uses a hidden state representation $h = (h_1, \ldots, h_T)$ so that it can map the input x to the output sequence $y = (y_1, \ldots, y_T)$. To compute this representation, it iterates through the following recurrence equations:

$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} h_{t-1}^{(i)} + \overrightarrow{b}^{(i)}) \tag{11}$$

$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} h_{t+1}^{(i)} + \overleftarrow{b}^{(i)}) \tag{12}$$

$$\hat{y}_t = g(Uh_t + c) = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c) \tag{13}$$

to show the input to each intermediate neuron at level i is the output of the RNN at layer $i-1$ at the same time-step t and output $\hat{y}$, at each time-step, is the result of propagating input parameters through all hidden layers (Eqn. 13).

To summarize, for a sequence of inputs $X = \{vc_t\}, t = 1, \ldots, T$ where, vc_t is one-hot encoded video id and context vector at time step t, EmoWare produces likely sequence of video to be used as recommendations, the joint probability $p(x)$ of which is represented as follows:

$$\begin{aligned} p(X) &= p(vc_1, \ldots, vc_T) \\ &= \prod_{t=1}^{t=T} p(vc_t | vc_{t-(n-1)}, \ldots\ldots, vc_{t-1}) \end{aligned} \tag{14}$$

i.e., we model p(X), the probability of the next video in a context given that we have a history of the sequence of videos with contexts.

IV. EXPERIMENTS

To investigate the performance of proposed EmoWare architecture, we conducted the experiment for a month and obtained results through 7 studies with different sets of 5 real users, ranging in age from 18-30 years having different educational backgrounds. Participants were provided with the participation agreements to obtain their consent for this research. Procured results are compared using various evaluation metrics with detailed description given in the following sections.

A. EXPERIMENTAL DESIGN

To implement the proposed EmoWare's system architecture, we created an end-to-end framework including a web application and a back-end model training server. The interface of the application presented to the users is shown in Fig. 5. Since collecting user opinion/feedback upon the recommendation of videos through manual mechanisms can be quite irritating, to automatize the feedback process without affecting and notifying the user, EmoWare's system is built with two monitors: 1) A video player and 2) A live camera stream to feed data of 33 facial points to Affectiva models. A real-time graph is provided for user reference, to display the intensity of induced viewer emotions as seen by the system on a continuous time scale. A detailed functionality of each of the components is described as follows:

- *Video Player:* One of the system's components which apart from playing the content, is an important source to collect user response. It provides the user, a choice to choose between movies and video songs through filters present in the extended control panel. Every user action such as skipping, liking or disliking, rating, forwarding, replying, etc., is constantly monitored

FIGURE 6. Examples of affective states of a viewer being captured by the system in the form of emotion-vector as the viewer watches video content.

and audited by the system. Each action has a reward associated with it which drives the system to make decisions about the user's present emotional preferences.

- *Live Web Camera:* We take into account the fact that analyzing user's mood, cognitive behavior in real time and the extent to which the user is getting affected while watching a video, can greatly aid in modeling temporal user preferences effectively. To achieve this, the system is built with a live webcam feedback mechanism to stream user's facial emotions and expressions data throughout the video length. Fig. 6 shows the affective feedback captured by the system as the viewer watches videos. As different viewers watch a video, the approach helps in evolving the initial annotated video corpus into a more precise representation of affective states and intensities to create a video timeline with general trends for the considered set of emotions. The general trends are then compared with the viewer's trend, to calculate the impact and take decision for the next recommendation, as described in section III-C.2.

B. DATASET DESCRIPTION

1) ANNOTATION

Although there are many publicly available databases for emotion recognition in videos and related affective domain, none of them provides continuous annotations based on the induced basic emotions in viewers for a sufficiently large length of movies as well as music videos. Generally, the available annotated datasets are only around valence-arousal dimensions. Some of them have global discrete labels on emotions while others have short-length videos (5 sec - 2 min long). The content of some datasets is either old and of very less interest to the user or they have been watched. As a result, the intensity of evoked emotions in viewers is less intense, mostly neutral rather. Also, the presence of both music and movie clips in the same database is a dearth.

The semantics of exhibited emotions is strongly related to the context of the situation and generally has cascading effects i.e. an emotion is an outcome of a series of activities happened in the past and not just instantaneous. Annotations provided by human labelers can best describe the impact and the intensity of induced emotions in such situations, thus, overcoming the limitation of machine-based

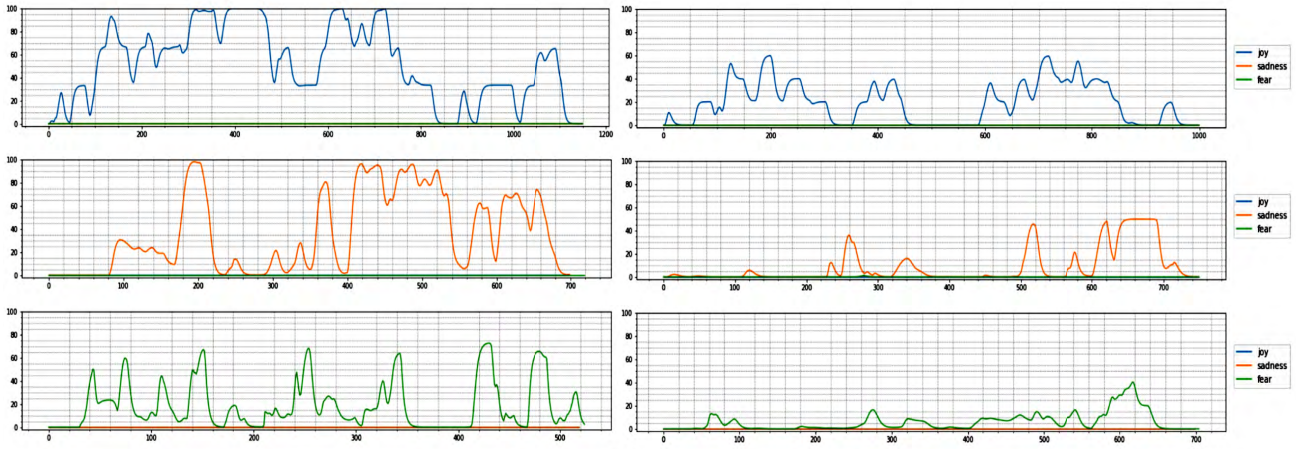

FIGURE 7. General affective intensity distribution graphs of 6 sample videos being obtained as the result of the annotation process. Graphs of the sample videos in the left and the right columns belong to high and low affective intensity categories respectively.

TABLE 4. Dataset description w.r.t quadrants of V-A Space.

Quadrant	I (+,+)	II (-,+)	III (-,-)
Songs	230	-	206
Movies	138	105	122

annotations where tracking context in videos is a very challenging task. That said, along with the best possible affective information extraction from FilmStim [29] and LIRIS-ACCEDE [22] movie databases, we manually collected videos from one of the most popular video sharing websites, YouTube. The distribution of the final dataset w.r.t valence-arousal space and intensity categorization is shown in Table 4.

After obtaining dataset, the annotation process is carried out in two phases. In the initial phase, it is achieved by extracting out key video frames using standard histogram color difference techniques [23], [24] and feeding them into Affectiva. This is done to seed the categorization process. The next level of annotation is achieved through crowdsourcing where more than 30 annotators participated. Since the degree of comfortability and ambiance affect one's mood, each one of them was required to simply explore the corpus as much as possible for a month with their webcam switched on to collect facial data while sitting in their comfort zone with no lab and experimental conditions. In order to get a naive feedback, only the first two feedback of a user per video is considered if the user repeats the visited/watched video. A live camera stream feeds the system with the facial points and Affectiva models provide continuous annotations based on streaming inputs. Furthermore, to validate the feedback, annotators are presented with a questionnaire at the end or before switching each video, containing questions related to their level of involvement, the perception of the content, context understanding, etc. to get precise information about the opinion of the user regarding the affective state of the watched content. As shown in [67]–[70] this is important because there can be differences in presented expressions v/s true perceived emotions and the estimates should reflect the true feelings of the annotator. Expected emotions are derived from the obtained answers using likelihood and correlation based techniques similar to [62]. Samples are then categorized into high, moderate and low affective intensity using the obtained continuous data and derived results above. The process results in a dataset containing videos in each category of emotion having following attributes:

- Video Id: A unique video identifier
- Video Name and Description
- Emotion Category: Joy, Sadness, Fear
- Affective Intensity Category: High, Mid, Low
- Intensity Based Ranking: Ranking in each category
- Intensity distribution along time axis
- Video Type: Movie clip, music video
- Language: English, Hindi

Fig. 7 represents the semantics of exhibited affective intensity distribution of 6 sample videos belonging to joy, sadness, and fear states respectively, obtained after the annotation process. Sample videos' affective curves presented at the left belong to high emotional intensities while those at the right belong to low emotional intensities. The intensity values are represented by ordinate axis w.r.t. a continuous time scale of 0.5 seconds, represented by abscissa axis.

2) INTER-RATER RELIABILITY

Inter-rater reliability is a measure to express the degree of agreement amongst the annotators taking part in an experiment. It is important to assess the consistency of annotations as it reflects the extent to which the data and information collected in a study are precise representations of the involved variables despite the subjective nature of a task. Percent agreement, Fleiss' kappa [58] and Krippendorffs alpha [59]

TABLE 5. Inter-annotator reliability per affective state.

Measure	Affective States		
	Joy	Sadness	Fear
Fleiss' κ	0.32	0.29	0.36
Krippendoff's α	0.21	0.19	0.15

are among the several statistical methods to measure inter-rater reliability but since percent agreement overestimates inter-annotator reliability as it does not take into account the agreement expected by chance, we, therefore, calculate Fleiss' kappa and Krippendoff's alpha to measure the annotation reliability. Their values range from -1 to 1, where, a value below 0 indicates disagreement among raters and negatively exceed what can be expected randomly, a value equal to 0 indicates no reliability, and a value higher than 0 represents an agreement between annotators (1 being perfectly reliable). Table 5 shows the analysis where it can be seen that all the values are positive and agreement is fair as per [61], implying agreement to be better than coincidence and is similar to various other studies [66] and [22].

C. EXPERIMENTAL SETTINGS

System initialization with a seed video of a particular emotion class can either be automatic or manual. The automatic feature allows a user to sit back and the system takes input from the facial data to identify the emotional state and starts playing accordingly. Otherwise, the user can manually select the required video from the list of videos under different emotions category. The next recommendation will happen dynamically based on the affective intensity of the present video and user actions (like fast forwarding, replying, skipping, rating, liking/disliking) towards the played video. The logged actions and user's facial analysis are used to make decisions by the system about the user preferences and to identify the video segments with high emotion intensities respectively. We define a session or an episode as the duration for which users watch videos, interact with the system until they log out. We split the dataset randomly into training and test sets consisting of 70% and 30% of the videos for each emotion category. To train the system, initially, sessions are generated with a set of 20 videos with varying emotions and intensities and optimizations are done using RL on streaming data. After certain iterations, enough data is recorded by the system to train DBRNN on it. Context-aware sequences are then generated by DBRNN and fed to RL for optimization. Eventually, we compute the average performance on each metrics as highlighted in section IV-E. We consider that feedback through video player have an edge over facial feedback through the camera, α and β are tuned to give slightly more weights to $R_{1(t)}$ and $R_{2(t)}$. Hence 0.34, 0.34 and 0.32 are chosen as corresponding weights of $R_{1(t)}$, $R_{2(t)}$ and $R_{3(t)}$ respectively.

1) SAMPLING RATE

It is worth mentioning that the efficiency and performance of the system are largely correlated as the sampling rate is largely affected by the hardware available. A high sampling rate of 30 FPS is usually preferred in order to properly capture certain categories of emotions which are highly spontaneous and transient such as fear. While desktop/laptop devices are capable of supporting such a high rate, the system is tuned to have a low rate of 5-15 FPS on resource intensive devices such as mobiles and tablets.

2) NETWORK TRANING DETAILS

EmoWare's neural network is trained using google's TensorFlow [13] with backpropagation through time. To determine the loss of trained model, we have used softmax-cross-entropy function (as it's a numerically stable function which internally computes the softmax activation). We tested the performance with RMSProp, SGD, and Adam Optimizers to optimize the loss function and chose later one as it was performing relatively better. Other hyper-parameters comprise LSTM cells of 512 hidden units, square root decay of learning rate from 0.01 to 0.001 as training proceeds, 40K training iterations ($\sim$ 40 epochs), gradient clipping of each cell using global norm with max_gradient norm is set to a max value of 10 to avoid gradient explosion.

D. SEARCH STRATEGY CLASSIFICATION

Search strategy or search behavior of users is highly dependent on the domain under study. Eminent studies and research in learning user models for text or music recommendation have benchmarks based on two general search strategy classes (exploration and exploitation) [2], [30]. Therefore, in this study, we classify users' search behavior into 2 broad categories as described below:

1) *Exploitative Users (Type-I):* are the ones who tend to remain in the same state of emotions and show opposing behavior upon recommending videos of different emotion class.
2) *Exploratory Users (Type-II):* are those who prefer to visit other class of emotions as well with non-negative feedback. To eliminate randomness and ensure a smooth transition from one emotion class to another, we recommend videos 1-ϵ times from the same class and ϵ times from other classes. Also taking into account the intensity factor, we recommend videos among the 3 videos which are closest to the E-Vector of the present video being watched using cosine similarity measure.

Each behavior is evaluated on a number of standard protocols, the details of which are given in section IV-E.

E. EVALUATION

To quantify EmoWare's system performance, we used five state-of-the-art evaluation metrics [9], [11]. We consider negative feedback as the error in prediction, observe the effect of top-k recommendations and false negatives (misses) on

user experience (which are more important than recall due to the stated reason), account for number of futile attempts to reach a satisfactory recommendation and the duration after which a user drops off or clicks away, henceforth, RMSE, Precision@K, Bounce Rate per episode, Miss-to-Hit Ratio and Average Watch-Time in our domain, are defined by as follows:

1) ROOT MEAN SQUARE ERROR PER EPISODE

One of the standard and pure academic accuracy metric in the bulletins of recommendation system: Correct prediction, often quantified in terms of RMSE, is defined as follows:

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(NF(i))^2} \tag{15}$$

where, *NF(i)* is the negative feedback received upon recommending i^{th} video and n is the total number of visited (watched) videos in an episode.

2) PRECISION@K

To measure the average proportion of top-k relevant recommendations, mean precision at k is defined as:

$$P@K = \frac{1}{N}\sum_{i=1}^{N}\frac{rV(i)}{k} \tag{16}$$

where, *rV(i)* is the number of relevant videos in i^{th} episode assuming n to be the size of set of relevant videos generated for an episode and N is the total number of episodes.

3) BOUNCE RATE (FNR OR 1-RECALL)

To measure the percentage of unsuccessful recommendations, bounce rate or false negative rate (FNR) is defined as:

$$Bounce\ Rate = \frac{1}{N}\sum_{i=1}^{N}\frac{drop(i)}{n} \tag{17}$$

where, *drop(i)* denotes the number of drops during i^{th} episode, n is the size of generated set for an episode and N is the total number of episodes.

4) MISS-TO-HIT (K)

Its a measures of number of unsuccessful attempts required to get k successful recommendations, defined as follows:

$$Miss - to - Hit(k) = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{drop(i)}{n - drop(i)}\right)k \tag{18}$$

where, *drop(i)* denotes the number of drops during i^{th} episode, n is the size of generated set for an episode and N is the total number of episodes.

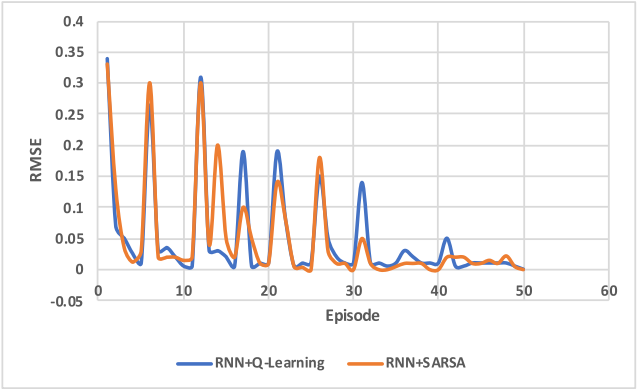

FIGURE 8. RMSE of exploitative user type.

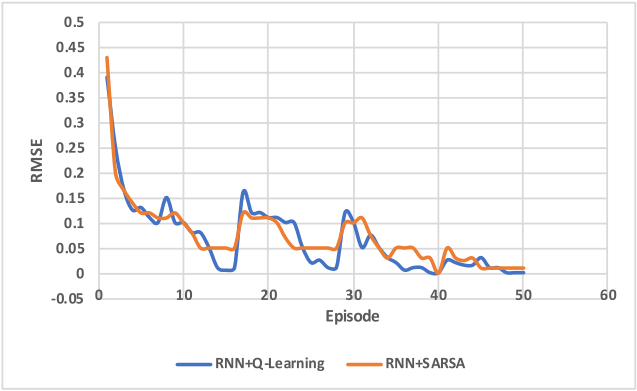

FIGURE 9. RMSE of exploratory user type.

5) WATCH-TIME

To account for user engagement over the total video length, average watch-time is defined as follows:

$$Watch - Time = \frac{1}{N}\sum_{i=1}^{n}\frac{WL(i)}{VL(i)} \tag{19}$$

where, n is the size of video set of an episode, *WL(i)* and *VL(i)* are the watched and total lengths of the i^{th} video respectively.

F. PERFORMANCE RESULTS

We recorded user feedback from multiple episodes (where each episode consisted of a set of 15 videos) and averaged out the results to see how much predictions are optimized. When a user starts interacting with the system with no prior knowledge, being an instance of a collaborative filtering recommendation system, it suffers from the cold start problem thus initial errors are high as shown in Figs. 8 and 9. But as the count of episodes increases, more user data is collected and we observe a large decrement in error.

In the case of the exploitative behavior of users, Fig. 8, it can clearly be seen that as the system encounters a new ecosystem, a surge in spike is witnessed, implying high negative feedback. But with increment in session count, there is a gradual decrease in height of such peaks and eventually, after 32 such sessions, both the algorithms converge to RMSE

TABLE 6. Performance for precision@k measurement.

User Type	Algorithm	P@5	P@10	P@15
Exploitative	RNN+SARSA	0.71	0.68	0.65
	RNN+Q-Learning	0.70	0.61	0.59
Exploratory	RNN+SARSA	0.80	0.65	0.54
	RNN+Q-Learning	0.81	0.78	0.69

TABLE 7. Performance statistics for RMSE and watch-time.

Metric	Measure	Shuffle	RNN+SARSA	RNN+Q-Learning
RMSE	Mean	9.81%	4.87% *	5.36% *
			7.26% **	6.41% **
	STD	11.2%	8.89%*	8.21%*
			6.80% **	7.21% **
Watch-Time	Mean	77.52%	83.93% *	80.06% *
			82.14% **	86.43% **
	STD	12.87%	8.93% *	7.03% *
			8.14% **	7.45% **

NOTE : * & ** represent User Type-I & II respectively.

less than 0.05. Fig. 9 shows the case of exploratory users' behavior in which, as compared to the previous case, both the algorithms take more number of sessions to converge to RMSE below 0.05. The behavior is quite expected as fuzziness makes it difficult to generalize patterns. But over time, the learning trend shows ideal behavior with decreasing errors.

If the two graphs of both user types are compared, it can be concluded that when users want to stay in the same emotional state (case I), spikes are more sudden and comparatively high, which leads to a conclusion that the cognitive state of exploitative users is more sensitive as compared to exploratory users and they resist affective migrations into other states.

Table 6 highlights the results for average precision @ 5, 10 and 15 respectively for all the possible user behavior and algorithm combinations. For P@5 almost both the algorithms perform similarly regardless of the search behavior of the users. But as K increases, RNN+Q-Learning starts showing clear dominance over RNN+SARSA for exploratory users while a total opposite behavior is observed in the case of exploitative users with an overall decrement in precision values. The behavior is quite expected as the exploration policy gets a larger search space to act with the increase is the size of the generated video subset.

Fig. 10 shows the observed bounce rate for Shuffle, RNN+SARSA and RNN+Q-Learning for both kinds of users. Here, we can clearly observe that RNN+SARSA outperformed shuffle by almost 27% in case of exploitative users while RNN+Q-Learning does it by almost 17% for exploratory users. Fig. 11 shows the results for getting 15 successful recommendations. Relative to shuffle, RNN+Q-Learning provides an edge of almost 5 videos for

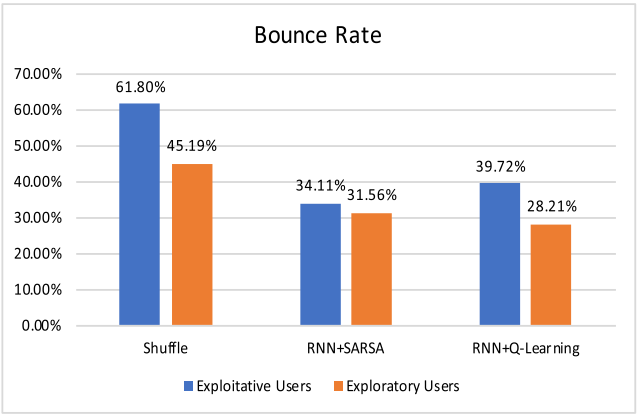

FIGURE 10. Bounce rate of exploitative and exploratory users.

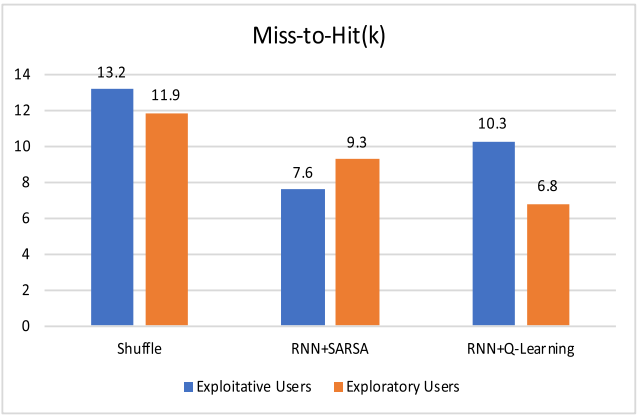

FIGURE 11. Miss-to-hit(k=15) for exploitative and exploratory users.

exploratory users while RNN+SARSA gives an advantage of almost 6 videos.

Table 7 list the mean and standard deviation of errors and watch-time for Shuffle, RNN+SARSA, and RNN+Q-Learning methods where the same behavior can be contemplated. We can observe that in the case of exploitative users, RNN+SARSA performs better than Shuffle and RNN+Q-Learning with less RMSE and more watch time while in case of exploratory users, RNN+Q-Learning outperforms the former two.

V. DISCUSSION

We compared our method with other affective analysis based video recommendation approaches which use facial expressions such as [78], [79]. Our hybrid algorithms generally outperformed these models, by having a much lower RMSE against the MAE results for different scenarios as shown in [78] and better computed emotions as shown in [79]. The results are quite apparent as we monitor and learn real-time contextual user feedback that too with multi-dimensional user behavior to recommend content based on the affective alignment of the user with the watched video and users' history pattern. We also compared our approach with other collaborative filtering and content-based video

recommendation models such as [71]–[74] which have been evaluated on Netflix and MovieLens datasets, where our method has shown noteworthy performance such as an average improvement of P@5 by 0.278. We also juxtaposed our approach with reinforcement learning based studies in music recommendations such as [2]. Since our model effectively models visual impact through videos which keep users more engaged with the process, it outperforms their model by showing improvements in almost all the metrics and certainly gives better comparison metrics such as decrement of 5.58% in bounce rate (using hybrid SARSA), difference of 2.6 in Miss-to-Hit(20) and 3.52% extra watch-time (using hybrid Q-Learning) etc.

As far as facial emotion recognition is considered, we used Affectiva's AFFDEX SDK which contains a data repository of 4 million faces analyzed in 75 countries. AFFDEX SDK performs face & facial landmark detection, face texture feature extraction, facial action classification and emotion expression modeling using both the standard and state-of-the-art techniques such as Voila-Jones face detection algorithm, the histogram of oriented gradient features, SVM for classification and other machine & deep learning techniques. It outperforms the existing state-of-the-art techniques as mentioned in [7]. To validate the performance of facial expression classification in our use-case, we have compared the results with the standard benchmarks using Cohn−Kanada database [77] where the percentage of accuracy improvement by our method is 1.13%, 15.8%, and 4.12% relative to [64], [75], [76] respectively.

Examining the limitations of the proposed technique, one of the major limitations is related to the relative ranking of the annotated video dataset which is based on the explicit information obtained from annotators and the implicitly derived results through interaction & facial affective data analysis. Alternatively stating, it is possible that videos at extreme ranks are not properly justified with their ranking. Defining the relative scale is also necessary as the annotation process ends up in one of the extreme direction (High, Mid, and Low) otherwise. Also, if not rated just after watching, the viewer might lose the affective impact and the context in which it was watched.

There are several other factors which potentially affect the annotators, the feedback process and the error in prediction.

Firstly, there is a heavy implicit reliability on the fact that data obtained from participants is close to a realistic representation of the induced emotions. It was impossible to ascertain that annotators had not done annotation basis perceived emotion or the emotion they thought they should have felt. There are works where it has been shown that perceived emotion differ significantly from annotation and rating [60].

Secondly, annotations made by participants are outside controlled lab conditions. There are numerous factors which can directly impact induced emotions in viewers including surroundings environment, the functionality of the gadget used and network connectivity. Regardless of these factors, inter-rater reliability reflects an overall agreement which could comparatively be better.

Thirdly, indeed, scalability is a major factor in the domain of affective computing, above shown results do not account for large-scale factors such as demography and geographic locations of the viewers, as shown in [63], where the impact of these factors is highlighted. The methodology has only been tested in a crowd composed of not very diverse ethnicity, interests, cultural backgrounds, educational backgrounds and age groups. It is one of the future work to test the efficiency of the algorithms in more difficult environments.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed EmoWare, a novel approach which combines the potential of reinforcement learning and deep bi-directional recurrent neural networks for automatic personalized video recommendation.

To address the problem of lack of affective information of a newly added video to a platform, such as YouTube, we have shown systematic techniques to process videos to get required affective intensity aware annotation to assimilate the emotion and affective intensity domain of the video to a certain extent which gets refined over time when multiple user feedback is collected. To capture user's non-verbal affective feedback towards a video, we track implicit as well as explicit user interactions along with facial expressions, using real-time video streams, throughout video length, to train personalized reinforcement learning models for learning short-term dynamic affective behavior. We also demonstrated how the sequence of videos being watched in different contexts can help in learning long-term affective trends by providing a rich-context aware feature set to a deep bi-directional recurrent neural network. The correlation is evaluated by conducting an experiment on two different video datasets with real users possessing different dynamic behaviors and obtained results are statistically compared to show the effectiveness of two algorithms in both the scenarios.

For further studies, we plan to evaluate the proposed approach on a more heterogeneous population with the increased dimensionality of the feature vector to mine user preferences more effectively. Taking audio and textual features of videos into account, a general mood-based multimedia content recommendation framework can be created which can monitor users' behavior through body language (e.g. gesture, posture, etc.) and not just facial expressions, in real time. Furthermore, our approach potentially opens the doors for other applications such as automatic video summarization which is worth studying too.

APPENDIX

ETHICS STATEMENT

DECLARATION

"Authors have obtained all ethical approvals from the subjects involved in this study".

ACKNOWLEDGMENTS

A. Tripathi was with the Department of Information Technology, National Institute of Technology Karnataka, Mangalore 575025, India. The authors thank all the participants and express their serious appreciation to the reviewers, editors, and colleagues from National Institute of Technology Karnataka, Surathkal, Mangalore, for providing insights and expertise during the course of this research.

REFERENCES

[1] A. Tripathi, T. S. Ashwin, and R. M. R. Guddeti, "A reinforcement learning and recurrent neural network based dynamic user modeling system," in *Proc. IEEE 18th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2018, pp. 411–415.

[2] C.-Y. Chi, R. T.-H. Tsai, J.-Y. Lai, and J. Y.-J. Hsu, "A reinforcement learning approach to emotion-based automatic playlist generation," in *Proc. IEEE Int. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2010, pp. 60–65.

[3] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 410–430, Oct./Dec. 2015.

[4] C. Simkins, C. Isbell, and N. Marquez, "Deriving behavior from personality: A reinforcement learning approach," in *Proc. Int. Conf. Cognit. Model.*, 2010, pp. 229–234.

[5] X. Y. Chen and Z. Segall, "XV-Pod: An emotion aware, affective mobile video player," in *Proc. IEEE World Congr. Comput. Sci. Inf. Eng. (WRI)*, vol. 3, Mar./Apr. 2009, pp. 277–281.

[6] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected 'in-the-wild,"' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 881–888.

[7] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behav. Res. Methods*, vol. 50, no. 4, pp. 1446–1460, 2018.

[8] *EMFACS Mappings Developed by Friesen & Ekman*. Accessed: Apr. 17, 2019. [Online]. Available: http://www.paulekman.com/product-category/facs/

[9] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 191–198.

[10] C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, 2016, Art. no. 13.

[11] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 2, pp. 95–116, 2018.

[12] K. Choi, G. Fazekas, and M. Sandler. (2016). "Towards playlist generation algorithms using rnns trained on within-track transitions." [Online]. Available: https://arxiv.org/abs/1606.02096

[13] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467

[14] M. Berglund, T. Raiko, M. Honkala, L. Kärkkäinen, A. Vetek, and J. T. Karhunen, "Bidirectional recurrent neural networks as generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 856–864.

[15] O. Olabiyi, E. Martinson, V. Chintalapudi, and R. Guo. (2017). "Driver action prediction using deep (bidirectional) recurrent neural network." [Online]. Available: https://arxiv.org/abs/1706.02257

[16] E. Smirnova and F. Vasile. (2017). "Contextual sequence modeling for recommendation with recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1706.07684

[17] X. Li *et al.* (2015). "Recurrent reinforcement learning: A hybrid approach." [Online]. Available: https://arxiv.org/abs/1509.03044

[18] J. A. Russell, "Affective space is bipolar," *J. Personality Social Psychol.*, vol. 37, no. 3, pp. 345–356, 1979.

[19] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.

[20] F. J. Massey, Jr., "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] 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.

[23] C. V. Sheena and N. K. Narayanan, "Key-frame extraction by analysis of histograms of video frames using statistical methods," *Procedia Comput. Sci.*, vol. 70 pp. 36–40, Jan. 2015.

[24] C. Sujatha and U. Mudenagudi, "A study on keyframe extraction methods for video summary," in *Proc. IEEE Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Oct. 2011, pp. 73–77.

[25] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[26] J. G. Ellis, W. S. Lin, C.-Y. Lin, and S.-F. Chang, "Predicting evoked emotions in video," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2014, pp. 287–294.

[27] IBM. (Mar. 30, 2017). *Tone Analyzer*. [Online]. Available: https://www.ibm.com/watson/developercloud/ tone-analyzer.html

[28] P. Boersma. (2006). *Praat: Doing Phonetics by Computer*. [Online]. Available: http://www.praat.org/

[29] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[30] M. Balabanović, "Exploring versus exploiting when learning user models for text recommendation," *User Model. User-Adapted Interact.*, vol. 8, nos. 1–2, pp. 71–102, 1998.

[31] R. Cowie, E. Douglas-Cowie, K. Karpouzis, G. Caridakis, M. Wallace, and S. Kollias, "Recognition of emotional states in natural human-computer interaction," in *Multimodal User Interfaces*. Berlin, Germany: Springer, 2008, pp. 119–153.

[32] M. Scheutz *et al.*, "Toward affective cognitive robots for human-robot interaction," in *Proc. Nat. Conf. Artif. Intell.* Cambridge, MA, USA: MIT Press, 2005, pp. 61–66.

[33] M. J. Duque, C. Turla, and L. Evangelista, "Effects of emotional state on decision making time," *Procedia-Social Behav. Sci.*, vol. 97, pp. 137–146, Nov. 2013.

[34] D. A. Worthy, K. A. Byrne, and S. Fields, "Effects of emotion on prospection during decision-making," *Frontiers Psychol.*, vol. 5, p. 591, Jun. 2014.

[35] J. Etkin, and A. P. Ghosh, "When being in a positive mood increases choice deferral," *J. Consum. Res.*, vol. 45, no. 1, pp. 208–225, 2017.

[36] A. Ogino and Y. Yamashita, "Emotion-based music information retrieval using lyrics," in *Proc. IFIP Int. Conf. Comput. Inf. Syst. Ind. Manage.*, Cham, Switzerland: Springer, 2015, pp. 613–622.

[37] J.-C. Wang, Y.-H. Yang, and H.-M. Wang. (2015). "Affective music information retrieval." [Online]. Available: https://arxiv.org/abs/1502.05131

[38] Y. Kim, Y. Shin, S.-J. Kim, E. Y. Kim, and H. Shin, "EBIR: Emotion-based image retrieval," in *IEEE Dig. Tech. Papers Int. Conf. Consum. Electron. (ICCE)*, Jan. 2009, pp. 1–2.

[39] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, Jan. 2009, Art. no. 421425. doi: 10.1155/2009/421425.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning—An Introduction*. vol. 1, no. 1. Cambridge, MA, USA: MIT Press, 1998.

[41] *Temporal Difference Learning Encyclopedia*. Accessed: Apr. 17, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Temporal_difference_learning

[42] G. Reynolds, D. Barry, T. Burke, and E. Coyle, "Towards a personal automatic music playlist generation alogorithm: The need for contextual information," in *Proc. 2nd. Audio Mostly Conf., Interact. Sound*. Limenau, Germany: Fraunhofer Institute for Digital Media Technology, 2007, pp. 84–89. [Online]. Available: https://arrow.dit.ie/cgi/viewcontent.cgi?article=1005&context=argcon

[43] Y. Hu and M. Ogihara, "NextOne player: A music recommendation system based on user behavior," in *Proc. ISMIR*, vol. 11, 2011, pp. 103–108.

[44] J. King and V. Imbrasaitė, "Generating music playlists with hierarchical clustering and Q-learning," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2015, pp. 315–326.

[45] M. B. Mariappan, M. Suk, and B. Prabhakaran, "Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2012, pp. 84–87.

[46] L. Chiarandini, M. Zanoni, and A. Sarti, "A system for dynamic playlist generation driven by multimodal control signals and descriptors," in *Proc. IEEE 13th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2011, pp. 1–6.

[47] L. Cardoso, R. Panda, and R. P. Paiva, "MOODetector: A prototype software tool for mood-based playlist generation," in *Proc. Simposio Inform.-INForum*, vol. 124, 2011. [Online]. Available: https://www.cisuc.uc.pt/publication/show/2635
[48] H. Yin, B. Cui, L. Chen, Z. Hu, and X. Zhou, "Dynamic user modeling in social media systems," *ACM Trans. Inf. Syst.*, vol. 33, no. 3, 2015, Art. no. 10.
[49] L. Ardissono and P. Torasso, "Dynamic user modeling in a Web store shell," in *Proc. 14th Eur. Conf. Artif. Intell.* Amsterdam, The Netherlands: IOS Press, 2000, pp. 621–625.
[50] A. Hawalah and M. Fasli, "A multi-agent system using ontological user profiles for dynamic user modelling," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 1, Aug. 2011, pp. 430–437.
[51] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1419–1428.
[52] T. Donkers, B. Loepp, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 152–160.
[53] Y. J. Ko, L. Maystre, and M. Grossglauser, "Collaborative recurrent neural networks for dynamic recommender systems," in *Proc. J. Mach. Learn. Res., Workshop Conf. Proc.*, vol. 63, 2016, pp. 366–381.
[54] x. Zhao, L. Zhang, Z. Ding, L. Xia, J. Tang, and D. Yin. (2018). "Recommendations with negative feedback via pairwise deep reinforcement learning." [Online]. Available: https://arxiv.org/abs/1802.06501
[55] X. Wang, Y. Wang, D. Hsu, and Y. Wang, "Exploration in interactive personalized music recommendation: A reinforcement learning approach," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1, 2014, Art. no. 7.
[56] O. C. Meyers, "A mood-based music classification and exploration system," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.
[57] M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi, "Personalizing session-based recommendations with hierarchical recurrent neural networks," in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 130–137.
[58] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, no. 3, pp. 613–619, 1973.
[59] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educ. Psychol. Meas.*, vol. 30, no. 1, pp. 61–70, 1970.
[60] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, p. 494, 2008.
[61] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[62] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Nov. 2005, pp. 381–385.
[63] R. Swinton and R. El Kaliouby, "Measuring emotions through a mobile device across borders, ages, genders and more," in *Proc. ESOMAR Congr.*, Atlanta, GA, USA, 2012, pp. 1–12.
[64] S. Zhao, H. Yao, and X. Sun, "Video classification and recommendation based on affective analysis of viewers," *Neurocomputing*, vol. 119, pp. 101–110, Nov. 2013.
[65] 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.
[66] S. M. Mohammad and D. P. Turney, "Crowdsourcing a word–emotion association lexicoN," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.
[67] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 47–56.
[68] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 526–540, Oct./Dec. 2016.
[69] S. Zhao, G. Ding, Y. Gao, and J. Han, "Learning visual emotion distributions via multi-modal features fusion," in *Proc. ACM Multimedia Conf.*, 2017, pp. 369–377.
[70] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," *Transfer*, vol. 1000, no. 1, pp. 4669–4675, 2017.
[71] Q. Zhu, M.-L. Shyu, and H. Wang, "VideoTopic: Content-based video recommendation using a topic model," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2013, pp. 219–222.
[72] C. Hongliang and Q. Xiaona, "The video recommendation system based on DBN," in *Proc. IEEE Int. Conf. Comput. Inf. Technol., Ubiquitous Comput. Commun., Dependable, Auton. Secure Comput., Pervasive Intell. Comput. (CIT/IUCC/DASC/PICOM)*, Oct. 2015, pp. 1016–1021.
[73] C. L. S. Bocanegra, J. L. S. Ramos, C. Rizo, A. Civit, and L. Fernandez-Luque, "HealthRecSys: A semantic content-based recommender system to complement health videos," *BMC Med. Inform. Decis. Making*, vol. 17, no. 1, p. 63, 2017.
[74] Q. Lu, T. Chen, W. Zhang, D. Yang, and Y. Yu, "Serendipitous personalized ranking for top-n recommendation," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 1, Dec. 2012, pp. 258–265.
[75] L. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, pp. 160–187, Jul. 2003.
[76] 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.
[77] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.
[78] I. Y. Choi, M. G. Oh, J. K. Kim, and Y. U. Ryu, "Collaborative filtering with facial expressions for online video recommendation," *Int. J. Inf. Manage.*, vol. 36, no. 3, pp. 397–402, 2016.
[79] Y. Diaz, C. O. Alm, I. Nwogu, and R. Bailey, "Towards an affective video recommendation system," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 137–142.

ABHISHEK TRIPATHI received the B.Tech. degree in information technology from the National Institute of Technology Karnataka, Mangalore, India, in 2017. Then, he joined Rivigo Services Pvt., Ltd., India, as a Software Development and Algorithm Engineer, in 2017. He has more than three research publications in reputed and peer-reviewed international conferences and journals related to reinforcement learning, deep learning, and artificial intelligence. His research interests include deep learning, affective computing, social network analysis, machine intelligence, and operational research. He is a Student Member of the IEEE.

T. S. ASHWIN received the B.E. degree from Visveswaraya Technological University, Belgaum, India, in 2011, and the M.Tech. degree from Manipal University, Manipal, India, in 2013. He is currently pursuing the Ph.D. degree with the National Institute of Technology Karnataka, Mangalore, India. He has more than 23 research publications in reputed and peer-reviewed international conferences and journal publications. His research interests include multi-modal affective content analysis, emotional, behavior and cognitive student engagement analysis, recommender systems, auto tutors, game-based learning, smart classrooms environments, and computer vision. He is a Student Member of the IEEE and ACM.

RAM MOHANA REDDY GUDDETI received the B.Tech. degree from Sri Venkateswara University, Tirupati, India, in 1987, the M.Tech. degree from the IIT Kharagpur, Kharagpur, India, in 1993, and the Ph.D. degree from The University of Edinburgh, U.K., in 2005. He is currently a Professor and the Head of the Department of Information Technology, National Institute of Technology Karnataka, Mangalore, India. He has more than 200 research publications in reputed and peer-reviewed international journals, conference proceedings, and book chapters. His research interests include affective computing, big data and cognitive analytics, bio-inspired cloud and green computing, the Internet of Things and smart sensor networks, social multimedia, and social network analysis. He is a Senior Member of the IEEE and ACM, a Life Fellow of IETE (India), a Life Member of the Computer Society of India, and a Life Member of ISTE (India).