Personalized Emotion-Aware Video Streaming for the Elderly

Yi Dong[1], Han Hu[2], Yonggang Wen[2], Han Yu[1], and Chunyan Miao[1,2]

[1] NTU-UBC Research Center of Excellence in Active Living for the Elderly, IGS, Nanyang Technological University
[2] School of Computer Science and Engineering, Nanyang Technological University
{ydong004,hhu,ygwen,han.yu,ascymiao}@ntu.edu.sg

Abstract. We consider the problem of video therapy services for the elderly based on their current emotional status. Given long hours watching TV in the elder population, most of the existing TV services are not geared for them. The elderly cannot tolerate complexity and negativity due to decline in cognitive abilities. In addition, the program is not adapted to the user's current emotional status. As a result, existing TV services can not achieve optimal performance across a broad set of user types and context. To provide content tailored to individual needs, and interests of the elderly, caregivers have to select an appropriate program manually. However, this can not scale well due to shortage of caregivers and high monetary cost. We present the personalized emotion-aware video streaming system, a redesign of conventional TV system to provide appropriate program flexibly, efficiently and responsively. Our proposed architecture adds video affective profiling, real-time emotion detection and Markov decision process based video program generation to the streaming service to this end. We present a complete implementation of our design. Trace-driven simulation has shown the effectiveness of our system.

1 Introduction

Watching television is a common leisure activity for older people. Retirees spend more than half their leisure time watching television according to American time use survey [15]. Many of them follow the broadcast schedule passively, which may contribute to the development of cognitive impairment. However, watching television can benefit the elderly in multiple ways if the program is selected carefully. Many studies have shown that watching TV can be mentally stimulating, thus benefit the elderly by anxiety and stress reduction, improved cooperative behavior and lowered use of psychoactive medications. Therefore, this type of non-pharmacological therapy video service can serve as multi-sensory stimulation and therapeutic use of music or video, which is recommended by many guidelines for the elderly, especially dementia caring [2].

The goal of this paper is to develop a non-pharmacological therapy video service. Building such a service is challenging on two fronts. First, video contents

should be cognitively congruent to the viewer's abilities, needs, and interests. Second, the viewer's needs changes dynamically. Thus, we need a system that is

1. **personalized** enough to meet the needs for each individual viewer.
2. capable of selecting video contents based on viewer's emotional status **in near real-time.**

Table 1. A summary of existing proposal and how they fall short of our requirements.

	Personalized	**Adaptive in real-time**	**cost-efficient**
Broadcast schedule	✗	✗	✓
VoD services	✓	✗	✓
Carehome staff guided	✓	✓	✗
Our approach	✓	✓	✓

As Table 1 shows, existing proposals fail on one or more of these requirements. For instance, broadcast schedules are not tailored for individual needs and adaptive to viewer's current emotional status. Similarly, Video on Demand(VoD) services provide control of the contents but they are not adapted as well. To address the aforementioned problems, traditional care-home relies on staffs to stay with residents so that they can choose the appropriate programs wisely and create a stimulus for interaction and conversation. Therefore, staffs themselves may need to monitor the emotional status of the elderly and find the signals when programs agitate the elderly. By long time observation, they may summarize the personal preference of each elderly and provide more accurate control. However, this kind of service cannot scale well. The shortage of care giver is becoming a big challenge and the monetary cost is a heavy burden. Thus, potential benefits promised by non-pharmacological video services remain unrealized.

We address these challenges and present the design and implementation of personalized emotion-aware video streaming system for the elderly. The design of our system is composed of accurate emotion detection using user equipment and the online personalized video program selection.

The user's emotion response can be measured accurately while watching the video by front cameras and wearable devices. We jointly employ the visual cues (facial expressions) and physiological signals like ECG to infer the emotional status of the viewer. Collectively, they can capture both the outer expressions as well as the inner feelings. We assume that users watch videos on devices with front cameras. Therefore, their frontal image with facial expressions can be easily captured. As for the ECG signals, a wrist band like Apple watch or Microsoft Band can measure accurately. Therefore, the emotion can be recognized and represented as a 2D emotion model [11] whose axes are valence and arousal.

To adapt video programs efficiently, we select video clips based on user's current emotional status. Our solution to content adaptation lies in the observation that the content characteristics are highly related to the emotion aroused [17]. Previous studies have developed affective models for video clips based on its

content characteristics. Therefore, we can choose the next clip with appro-priate affective properties based on user's current emotional status, optimizing the overall experience.

Our goal is to select appropriate video based on user's current emotion. Specifically, the following goals need to be achieved.

- Stable arousal. Read the signals when current video agitates someone and avoid this type of videos.
- High valence. Find out what the viewer really likes, and achieve the maximum variance of content and emotional experience.
- Exploration and exploitation tradeoff. To avoid boredom, the viewer should be exposed to new video clips with different ideas, art and culture. However, the affective effects of new clips are not verified.

The reward function should balance the three components.

In this paper, we demonstrate that, for such decision making, Markov decision process (MDP) can derive the optimal policy if the emotion-video model is known.

We implemented a prototype of non-pharmacological therapy video service. In real trace-driven evaluation, we assess our proposed algorithms based on publicly available emotional response dataset `DEAP`. The experiments show that our system is usable and robust.

We summarize our key contributions as follows.

- Design of system architecture for the personalized emotion aware video streaming system (§3).
- Formulating the video program adaptation as an Markovian Decision Process (§4).
- Real-world implementation of the system and trace-driven evaluation. (§5).

2 Related Works

Our system is built based on previous research on affective video content modeling, multimodal user emotion detection, adaptive video streaming. However, our system differs from previous work from multiple perspectives.

Adaptive video streaming: As a killer application of the Internet, video streaming has received tremendous attention in both industry and academia [12]. Nowadays, adaptive video streaming services are usually hosted by cloud computing platforms where video clips has been transcoded into different bitrate versions [5–7]. Existing client-side video players employ adaptive bitrate (ABR) algorithms to optimize user quality of experience (QoE). ABR algorithms dynamically choose the bitrate for each video segment based on user's network conditions or client buffer occupation [8]. This line of research focus on the resource allocation and optimization based on network resource and device usage [9]. In contrast, our proposal is the first effort on video adaptation based on viewer's current emotional status.

Video services for the elderly: The Gerontological Society of America

published Video Respite™ , which are video tapes developed for the persons with dementia. Early research shows the tapes to be calming for the persons with dementia [13]. Recent studies suggest that such type of video services has considerable promise as an individual and group activity in long-term setting [1]. `Memory-Lane.tv` [3] has developed an interactive and multi-sensory media collection for the memory impaired. It supports modern delivery methods such as tablets, computers and Apple TV. However, these video services require staffs to operate following guidelines to achieve promising results. Our system can release the burden of the caregivers and can be deployed easily.

3 System Architecture

In this section, we present our proposed system architecture for personalized emotion-aware video streaming systems. It is possible to adapt the program based on user's current emotional status. Specifically, we give detailed illustration on the three components: real-time personalized multimodal emotion detection, video profiling and MDP-based program adaptation.

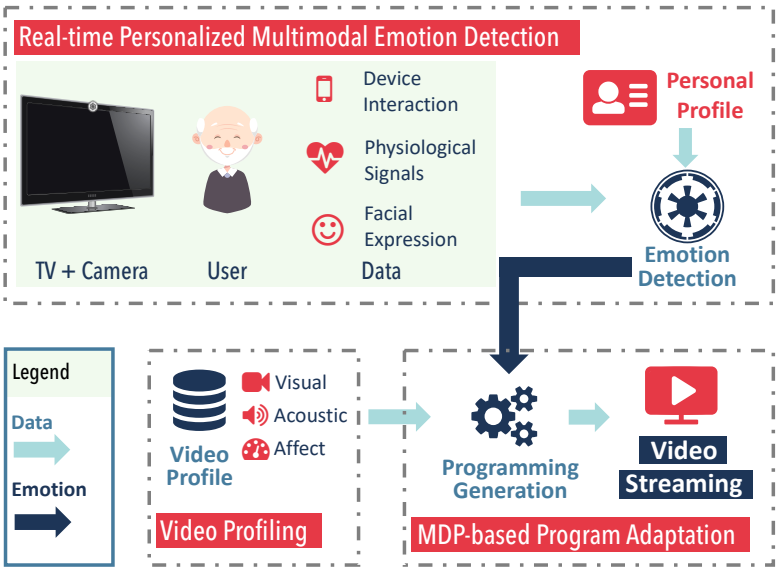

Fig. 1. An overview of the personalized emotion aware video streaming system. The system captures user's emotional status and adapts the program accordingly in a real-time manner.

[3] http://memory-lane.tv/

3.1 Emotion Detection.

Emotion response while watching video is being sensed by ECG (Electrocardiography) signals and facial expressions. These signals may complement each other. Facial expressions are captured using front cameras which are ubiquitous in mobile devices while ECG signals are sensed by smart wrist or extracted from facial images.

3.2 Video Affective Profiling.

This module can analyze the emotional impact of a video content will have on viewers, in terms of valence and arousal.

3.3 Video program selection.

A set of videos that can provide a variance of video play experience is stored in our server. The system should work in the similar way that staffs in care giving institutions do to optimize the user's experience. Our goal is to select appropriate video based on user's current emotion. Specifically, the following goals need to be achieved.

- Stable arousal. Read the signals when current video agitates someone and avoid this type of videos.
- High valence. Find out what the viewer really likes, and achieve the maximum variance of content and emotional experience.
- Exploration and exploitation tradeoff. To avoid boredom, the viewer should be exposed to new video clips with different ideas, art and culture. However, the affective effects of new clips are not verified.

The reward function should balance the three components.

4 Emotion-aware MDP-based Program Adaptation

we now give specific math models for each components of our system. Based on the models, we have formulated the operation of emotion aware video streaming as an MDP.

4.1 System Models & Assumptions

In this subsection, we present specific mathematical models for the emotion-aware content enhancement system, including user emotional status model, video response model and content caching model. For ease of reference, we summarized the notations in the Table 2.

Table 2. Definitions of Notations

Notation	Definition
s, s'	States
a	Action (The video to play)
r	Reward
$\mathcal{S}$	Set of all nonterminal states
$\mathcal{A}(s)$	Set of all actions possible in state s
$\mathcal{R}$	Set of all possible rewards
t	Discrete time step
T	Final time step of an episode
S_t	State at time t
A_t	Action at time t
R_t	Reward at time t
π	Policy, decision-making rule
$\pi(s)$	Action taken in state s under deterministic policy π
$\pi(a\|s)$	Probability of taking action a in state s under stochastic policy π
$p(s'\|s, a)$	Probability of transition to state s', from state s taking action a
γ	Discouont-rate parameter
$v_\pi(s)$	Value of state s under policy π (expected return)
$v_*(s)$	Value of state s under the optimal policy
V, V_t	Array estimates of state-value function v_π or v_*

Emotional Status Model As shown in Figure 2, we capture a person's emotions from both outward expressions and inner feelings. Specifically, we develop a machine learning based model to infer user's emotional status from physiological signals (eg. ECG) and facial expressions. We adopt a 2D emotional model [11] whose axes are valence (pleasure to displeasure) and arousal (high to low).

Video Response Model Videos can render the experience necessary to arouse an emotion. The dynamics of emotion is the hall mark of a successful video. Emotions are created in the film's text. Since emotion is a short-term experience that peaks and burns rapidly, we can use a transitional matrix $\mathcal{T}$ to capture the dynamics of emotions. In this matrix, $P_{sa} = P(s, s', a) = \Pr(s_{t+1} = s'|s_t = s, a_t = a)$ denotes the probability that video a in state s at time t will lead to state s' at time $t + 1$.

Accumulated Rewards Model The viewers are satisfied by the videos. We define the immediate reward received after the transition to a new emotion s' with

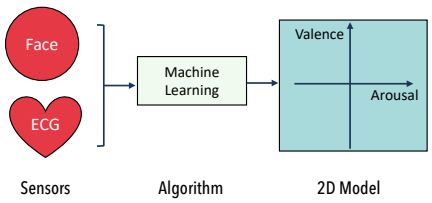

Fig. 2. The viewer's emotion can be extracted jointly from his/her facial expression and ECG signals. The Arousal-Valence model is used to model the emotion.

video a as $R_{sa} = R(s, s', a)$. We assume that viewers gain more rewards while watching more videos. Moreover, we consider the law of diminishing returns. The more we experience something, the less effect it has. Therefore, we put a discount factor $0 \leq \gamma < 1$ to capture this characteristic. Therefore, the accumulated reward over infinite horizon is $\sum_{t=0}^{\infty} \gamma^t R(s_t, s_{t+1}, \pi(s_t))$. Please note that if the status is in termination, the reward will be 0.

4.2 MDP Formulation

As shown in Figure 3, the viewer and TV interact in a sequence of discrete time steps, $t = 0, 1, 2, 3, \ldots, T$. At each time step t, the mobile device receives the representation of viewer's emotion state, $S_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible emotions and on that basis selects an *video*, $A_t \in \mathcal{A}(S_t)$, where $\mathcal{A}(S_t)$ is the set of possible states available in state S_t. On time step later, in part as the impact of the video viewed, the viewer's emotional needs are fulfilled and can be quantified as a numerical reward, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ and find oneself in a new emotional state, S_{t+1}. The figure above diagrams the viewer-TV interaction.

At each time step, the mobile device implements a mapping from states to probabilities of selecting each possible videos. This mapping is called policy and is denoted π_t, where $\pi_t(a|s)$ is the probability that $A_t = a$ if $S_t = s$.
In a real system deployment, we can model this problem as a Markov decision problem (MDP). An MDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}_{sa}, \mathcal{R}_{sa}, \gamma)$ as defined in the previous subsection. The solution of the MDP is a policy π that tells us how to make decision, i.e., choose an action when a particular state is observed. There are many possible policies but the goal is to derive the optimal policy π^*, which can be formally given as:

$$\pi^*(s,t) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left[\sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s',t)[\mathcal{R}(s,a,s',t) + V^*(s',t+1)] \right], \qquad (1)$$

where V denotes the value function. The optimal policy can be developed by dynamic programming algorithms [16].

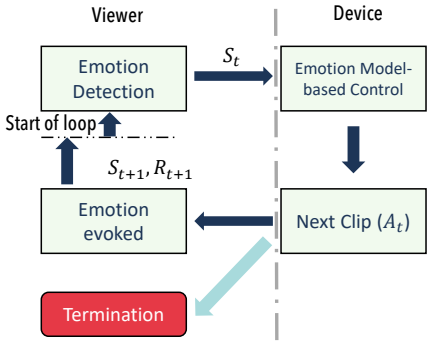

Fig. 3. The viewer-TV interaction in Markov Decision Process based program generation. The loop starts from the emotion detection.

5 System Implementation

We have built a prototype of personalized emotion-aware video therapy service that closely follows the architecture described in the previous sections. This prototype uses Microsoft band

5.1 Hardware Settings

For the client side, our system support most of the main stream mobile devices (e.g. Android/iOS) or PC that connected with a webcam. Users can log in our web portal (Figure 4) to check the status of the sensor connection. Once the user has authorized the web browser to use webcam are connected, the video session is ready to start.

In our current implementation, the client side captures the frontal face image via front cameras and transmit the compressed image to the server side for emotion detection. Under good lighting condition, both facial expressions and ECG signals are extracted from the webcam image. Instead of relying on a webcam to capture ECG signals, We recommend users to wear ECG sensors if the lighting condition is not satisfactory.

Multimodal emotion detection and video program adaptation is deployed in the server side, given high computational demands. Our server is with 2 Intel® Xeon® E5-2603, 2 Nvidia K80 graphic cards and 64GB memory.

5.2 Software Implementation

The user interface of the client front-end is shown in Figure 4. We have implemented a web service for video streaming and affective profile visualization. The video player on the left can enter full screen mode to provide an immersive experience. The affective profile is visualized in the line chart on the right side.

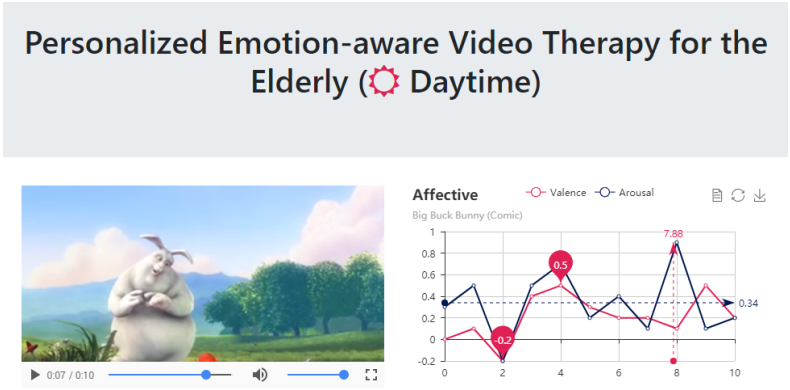

Fig. 4. Screenshot of the front end of personalized emotion-aware video service.

A vertical dotted line is synchronized with the player to indicate the current playing time.

We had implemented the affective video profiling and multimodal emotion recognition by machine learning algorithms. Specifically, we trained SVR (Support Vector Regression) models from `DEAP` dataset [10], using `LibSVM` [3]. Audio features have been extracted by `openSmile` toolbox [4]. In addition, general video features were extracted using `LIRE` library [14]. Beyond that, we have extracted CNN features using Matlab neural network toolbox. The models in this part were developed for proof of concept. We believe that the accuracy of the two models will increase if more advanced deep learning algorithms are employed.

For personalized emotion-aware video program adaptation, we implement the proposed MDP model using Markov Decision Process (MDP) Toolbox for Python[4].

5.3 Limitations

We acknowledge three potential limitations in our study.

- **Lighting condition and user's head position**. Accurate emotion detection requires clear frontal facial image. The motion of viewer's head will cause motion blur which is the noise for our emotion detection algorithm. Therefore, our system only works under good light condition where viewer's head seldom moves.
- **Delay due to computation and communication**. Given high computation demands and the large size of media files, the delay of emotion detection is noticeable. In real-world deployment, we should strike a good balance between accuracy and delay.

[4] `https://github.com/sawcordwell/pymdptoolbox`

– **Generalization of the emotion detection model**. In this paper, we have designed multimodal algorithms based on `DEAP` dataset. We have not tested the generalization of the emotion detection models in other settings. We plan to investigate this problem in our future work.

6 Conclusions

In this work, we designed and implemented a personalized emotion-aware video therapy system for the elderly. Unlike conventional TV systems that use fixed program, we adapt the program based on user's current emotional status and the video affective attributes. We have formulated the emotion-aware video program adaptation as an MDP problem which can be solved by dynamic programming algorithms. We have deployed the emotion detection, video profiling and video program adaptation on a server with Nvidia K80 GPU. The system can effectively adapt the video program based on user's current emotional status. Trace-driven simulation validates that it is a robust and usable system.

7 Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative; and the Interdisciplinary Graduate School, Nanyang Technological University, Singapore.

References

1. Caserta, M.S., Lund, D.A.: Video respite® in an alzheimer's care center: Group versus solitary viewing. Activities, Adaptation & Aging **27**(1) (2003) 13–28
2. Centre, N.C.: Dementia: A nice-scie guideline on supporting people with dementia and their carers in health and social care, British Psychological Society (2007)
3. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3) (2011) 27
4. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia. MM '13, New York, NY, USA, ACM (2013) 835–838
5. Gao, G., Hu, H., Wen, Y., Westphal, C.: Resource provisioning and profit maximization for transcoding in clouds: A two-timescale approach. IEEE Transactions on Multimedia **19**(4) (April 2017) 836–848
6. Gao, G., Wen, Y., Hu, H.: Qdlcoding: Qos-differentiated low-cost video encoding scheme for online video service. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications. (May 2017) 1–9
7. Gao, G., Zhang, W., Wen, Y., Wang, Z., Zhu, W.: Towards cost-efficient video transcoding in media cloud: Insights learned from user viewing patterns. IEEE Transactions on Multimedia **17**(8) (Aug 2015) 1286–1296
8. Hu, H., Wen, Y., Niyato, D.: Spectrum allocation and bitrate adjustment for mobile social video sharing: Potential game with online qos learning approach. IEEE Journal on Selected Areas in Communications **35**(4) (April 2017) 935–948

9. Hu, H., Jin, Y., Wen, Y., Chua, T.S., Li, X.: Toward a biometric-aware cloud service engine for multi-screen video applications. In: Proceedings of the 2014 ACM Conference on SIGCOMM. SIGCOMM '14, New York, NY, USA, ACM (2014) 581–582
10. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis ;using physiological signals. IEEE Transactions on Affective Computing **3**(1) (Jan 2012) 18–31
11. Lang, P.J.: The emotion probe: Studies of motivation and attention. American psychologist **50**(5) (1995) 372
12. Li, B., Wang, Z., Liu, J., Zhu, W.: Two decades of internet video streaming: A retrospective view. ACM transactions on multimedia computing, communications, and applications (TOMM) **9**(1s) (2013) 33
13. Lund, D.A., Hill, R.D., Caserta, M.S., Wright, S.D.: Video respite: An innovative resource for family, professional caregivers, and persons with dementia. The Gerontologist **35**(5) (1995) 683–687
14. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: Proceedings of the 16th ACM international conference on Multimedia, ACM (2008) 1085–1088
15. Robinson, J., Godbey, G.: Time for life: The surprising ways Americans use their time. Penn State Press (2010)
16. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction (2nd edition). Cambridge, MA: MIT Press (2017)
17. Zhu, Y., Hanjalic, A., Redi, J.A.: Qoe prediction for enriched assessment of individual video viewing experience. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM (2016) 801–810