Multimedia Tools and Applications (2021) 80:2887–2905
https://doi.org/10.1007/s11042-020-08836-3

Deep learning-based late fusion of multimodal information for emotion classification of music video

Yagya Raj Pandeya[1] · Joonwhoan Lee[1]

Received: 10 April 2019 / Revised: 17 February 2020 / Accepted: 9 March 2020 /
Published online: 17 September 2020

Abstract

Affective computing is an emerging area of research that aims to enable intelligent systems to recognize, feel, infer and interpret human emotions. The widely spread online and off-line music videos are one of the rich sources of human emotion analysis because it integrates the composer's internal feeling through song lyrics, musical instruments performance and visual expression. In general, the metadata which music video customers to choose a product includes high-level semantics like emotion so that automatic emotion analysis might be necessary. In this research area, however, the lack of a labeled dataset is a major problem. Therefore, we first construct a balanced music video emotion dataset including diversity of territory, language, culture and musical instruments. We test this dataset over four unimodal and four multimodal convolutional neural networks (CNN) of music and video. First, we separately fine-tuned each pre-trained unimodal CNN and test the performance on unseen data. In addition, we train a 1-dimensional CNN-based music emotion classifier with raw waveform input. The comparative analysis of each unimodal classifier over various optimizers is made to find the best model that can be integrate into a multimodal structure. The best unimodal modality is integrated with corresponding music and video network features for multimodal classifier. The multimodal structure integrates whole music video features and makes final classification with the SoftMax classifier by a late feature fusion strategy. All possible multimodal structures are also combined into one predictive model to get the overall prediction. All the proposed multimodal structure uses cross-validation to overcome the data scarcity problem (overfitting) at the decision level. The evaluation results using various metrics show a boost in the performance of the multimodal architectures compared to each unimodal emotion classifier. The predictive model by integration of all multimodal structure achieves 88.56% in accuracy, 0.88 in f1-score, and 0.987 in area under the curve (AUC) score. The result suggests human high-level emotions are automatically well classified in the proposed CNN-based multimodal networks, even though a small amount of labeled data samples is available for training.

Keywords Emotion classification · Music video dataset · CNN · Multimodal approach · Late fusion

✉ Joonwhoan Lee
chlee@chonbuk.ac.kr

Extended author information available on the last page of the article

1 Introduction

Music is a language that communicates some emotion to anyone, even to plants or animals, and visual perception is more playing a crucial role in our daily lives as they aid decision-making, learning, communication, and situation awareness in human-centric environments. The music artists generally use dynamics (tempo, meters) or articulation to evoke emotions in music, but the listeners can have different feelings and sentiments. In general, the human emotion incurred from visual or acoustic information is not only vague and subjective, but also depends on human thought and environmental changes. This kinds of vagueness usually reflecting on music video emotion analysis.

In our modern society, the music videos are one of the most floating contents over the Internet. The invention of social media sites and the commercial music management sites, the Internet becoming dense day by day with more user's demands. Today the customer's test for music video selection is not only limited to its name, album, and artist but moved towards more attributes like genre, mood, visual expression, and visual quality. Thus, an immense amount of music video needs to be classified according to these attributes, preferably in an automated way. This kind of need can be seen highly in online music video stores, music galleries, digital music market, and file-sharing networks. The proper evaluation can help the music manager to better understand the social demand and end user test.

Several methods are proposed using audio and video multimodal information for a very broad range of applications, but the music video emotion analysis is still an unsolved and challenging problem. Human emotion can be expressed either verbally through emotional vocabulary or by expressing nonverbal cues such as intonation of voice, facial expressions, and gestures. There have been successful research breakthroughs on emotion recognition based on stimuli including video, text, speech, facial expressions, gestures and eye gage for high-level semantic prediction. The emotion in music-video includes such emotional attributes and the additional sentiments express through music melody, instrumental rhythm, and composer highlighted scenes. In this research, we focus to classify the music video emotion by convolutional neural networks. To consider the proper structures of the emotion classifier network, we made a comparative analysis of recently well-known deep neural networks (DNNs) and fine-tune each of them with a small labeled dataset to adapt its characteristics. The unimodal DNNs taken each of the audio or video networks are integrated for four types of multimodal architectures. For better understanding, before discussing the proposed multimodal approach, it would be better to discuss some past frontline unimodal and multimodal approaches using video or(and) audio as input.

Video is a sequence of correlated pictures information widely exploited in diversified domain such as classification, object detection and tracking [64], 3D object retrieval [16], human action recognition [15], human emotion prediction [27, 36], object segmentation [65], autonomous system [14], and object pose estimation [59] in last decade. The popular video classification approaches include the frames majority vote [24], temporal feature pooling (TFP) [37], three-dimensional (3D) convolution (C3D) [58], Inception architecture with 3D convolution (I3D) [4], and recurrent neural networks (RNNs) [61] based method. The frames majority vote and TFP are image-based methods and do not work well for videos as they do not include temporal information. Even though RNNs and their variants such as long-short term memory (LSTMs) networks operate on the frame-level features captured from convolutional neural network (CNN) activations similar to TFP but the short time sequence modeling ability of RNNs is not enough to integrate whole video information. The C3D and

I3D are well- accepted structure that is amenable for spatiotemporal feature learning. These architectures are able to capture the whole video information and can make high level decision like emotion category.

Similar to visual emotion, music is another rich source of human emotion because it explains the artistic emotion through lyrics and musical instrument play. Many neural network models for music information retrieval (MIR) uses two-dimensional (2D) Mel spectrogram as input. The research [1, 6] shows the transfer learning provides better results for music genre classification. The CNN based method in [32] uses music spectrogram as input to predict music emotions. Instead of the spectrogram, however, CNN with raw waveform input [31] are also getting popular because the spectrogram is a handcrafted magnitude-only representation without phase information. We consider both these music data input methods in our proposed multimodal architectures.

The wide proliferation of videos posted online increasingly moves the active research direction from conventional unimodal to complex forms of multimodal approach. Emotion from multimodal information defines the presence of more than one modality or channel. Multimodal approach for music video emotion explores the feeling or sentiment of music composers included in song lyrics, musical instrument play, and visual expression. There are growing opportunities for automatic emotion recognition systems as the technology and the understanding of emotions are advancing. Most of the researches is intended to explore human emotion based on facial expression or human actions. The study [23] proposes CNN based architecture for facial emotion recognition of humans. An extension on face emotion analysis is proposed on [69] using an audio spectrogram and human face image based on an integrated multimodal architecture. The multimodal approaches [11, 13, 41, 44] have proposed audio and video by using a recurrent network with LSTM cells for face video emotion recognition. The one-dimensional (1D) audio network- and 2D video network-based multimodal [61] for speech recognition uses hybrid information fusion techniques by adding recurrent neural network after concatenation of learned features. The audio (2D or 3D) and video (3D) multimodal [34, 38, 49] with deep belief network (DBN) is proposed for face video emotion recognition. Besides the human facial emotion, movie action based human emotion analysis is performed in [56]. The study in [55] proposes audio, video, and text integrated multimodal architecture for natural disaster information management. Recent work in multimodal emotion fusion [40] integrates audio, video and text modalities for human emotion recognition. This study uses the RECOLA dataset [46] where participant's emotions have been recorded in dyads during a video conference. The review papers in [2, 45, 62] also include various affective computing based on the multimodal techniques.

Audio-video multimodal approach discussed above mainly focused either on human facial emotion or human action-based emotion, but the music video-based emotion is isolated from the current active areas of research. As far as we know, there is no deep learning-based music video emotion recognition research, except for some electroencephalography- (EEG) based study [47, 60]. In this research, we aim to recognize the emotion of music video and make a comparative study among various unimodal and multimodal structures. Our work exploits the music and video information only for music video emotion prediction. The primary limitation in this area of research is the lack of labeled data and hence we first introduce a small music video dataset. When the amount of training data is not enough, transfer learning is also a popular way to treat the data scarcity problem if the data domain of source and destination well matches in their characteristics. So, we adopt some pre-trained convolutional audio and video networks and fine-tune them to adapt the new music video environment. We also propose a 1D

CNN-based music network with raw waveform as input that can preserve the phase and magnitude information of music data. All unimodal performances are comparatively analyzed over various optimization techniques. The best outcomes of a pair of unimodal approaches are integrated on multimodal DNN using late fusion technique. Finally, these four music video multimodal are also integrated to get the final prediction together and evaluated over the various metric. All the proposed multimodal structure uses cross-validation to overcome the data scarcity problem at the decision level. In general, our main contribution in this study is a small music video dataset for human emotion analysis that can be an attraction point of new researchers in this research domain. The other focus of this study is an analysis of currently well-known unimodal CNN structures with various tuning parameter and their contribution to integrating with other unimodal architecture particularly for human emotion prediction. In this overall research, we found that the multimodal results show improvement in various evaluation matrices over unimodal performance.

The paper is organized as follows: In Section 2, we present the possible emotion state representations and explain our music video emotion dataset. Section 3 includes deep neural network architectures with their input processing and the learned feature fusion of music video emotion dataset. Section 4 illustrates the proposed unimodal and multimodal classification results based on various evaluation metrics. Finally, the conclusion of this study is included in Section 5.

2 Music video emotion dataset

Emotion is a psycho-physiological process triggered by conscious and/or unconscious perception of an object or situation and is often associated with mood, temperament, personality and disposition, and motivation. Emotion plays a vital role in human communication, decision handling, interaction, and cognitive process. Emotion is based on subjective experiences, and people represent them with many semantic terms. Many emotion representation models have been proposed in past decays. The categorical model divides emotion into several discrete classes and applies machine learning techniques to train a classifier. The dimensional model defines emotion as numerical values over several emotion dimensions (e.g., valence and arousal [48, 54]). These regression models are trained to predict the emotion values that represent the affective contents of input data, thereby representing the input as a point in the emotion space.

Although emotional experiences are represented within a semantic space that can be captured by categorical labels, the boundaries between categories are vague rather than crisp. Usually, the human emotion representation models are a little different according to the modalities to evoke emotion (for example, speech, music, video, facial expression, and real time action video). The study in [18, 66] provides a comprehensive review of the modalities that have been proposed for music emotion recognition. The speech emotion recognition study [39] illustrates six categories of human vocal emotions. Among various facial emotion studies, [5, 68] represent the facial expressions with seven categories in two-dimensional valence-arousal space. A three-dimensional model for eight basic facial emotions and monoamine neurotransmitters is represented in [33] using the corners of a cube.

The major challenges of music video emotion analysis are the vague emotion boundary, personality, and scarcity of labeled training data. The DEAP [29] database is the first publicly available database composed of only 120 one-minute-long excerpts of music videos. Each one was rated by at least 14 volunteers from an online self-assessment based on induced arousal, valence and dominance. The dataset cannot lead the present requirement of data to well train the data driven based algorithms. The data scarcity problem for data-driven emotion analysis in the case of the music video is a great challenge for researchers. The emotion is a high-level semantics of any information processing system and require more samples for more accurate decision. There is no alternative to solve the lack of data problem without increasing the data samples. Hence, we integrate some samples from DEAP dataset and a vital study [21] for music and video retrieval, and other samples are collected from the Internet. Finally, we introduce a small dataset for music video emotion analysis with a nearly equal number of samples in each emotion classes as illustrated in Table 1.

We integrate various human emotion adjectives into broad six categories as basic emotion classes in our proposed dataset. The study [8] represents 27 distinct possible categories of human emotion but in case of music video, it is convenient to organize them with coarse semantic groups so that an end-user can easily demand the required music video from large video banks or online music video stores. We categorize the adjectives of music video emotion classification into six basic emotion categories with references [41, 52, 67], namely, *Exciting*, *Fear*, *Neutral*, *Relaxation*, *Sad*, and *Tension*. From each emotion class, respectively three samples are represented (from left to right) in Fig. 1 and each music video sample is approximately 30 s in length.

In this music video emotion dataset, most music videos are collected from the Internet and this makes huge diversity in our dataset in terms of territory, languages, cultures, and musical instruments. Each data sample is distinct by their various features including frequency, pitch, energy, zero-crossing rate, motion intensity, color energy, lighting, rhythm regularity, frames, resolutions, etc. All these factors play a vital role to make each human emotion distinct which is express through facial expressions, eye states, mouth movements, and body actions. This is the reason that makes the emotion analysis highly confusing and nuanced. The vague boundary of emotion classes and their correlation to each other is shown in Fig. 2. We have selected music videos which give temporarily consistent and easily determinable emotions, and annotated each of them with one of the adjectives corresponding to 5 basic categories. The *Exciting* state of human emotion includes positive or pleasant emotions generally seen in the context of human satisfaction, subjective well-being, and placement. The visual contents generally include high body movements such as dance or party, group activity, and colorful environment. This type of music general adopts fast tempo, major tonality, complementing harmonics and smooth or varied rhythm. The *Fear* emotion arises from the perception of danger or horror. The visual information includes some unnatural events or characters that change abruptly in time. The high tension appears in human expression and action. The music

Table 1 Music-video dataset with various adjectives and number of data samples in each emotion classes

Emotion class	Emotion adjectives	No. of samples
Excitation	Happy, Fun, Love, Sexy, Joy, Pleasure, Exciting, Adorable, Cheerful, Surprising, Interest	720
Fear	Horror, Fear, Scary, Disgust, Terror	519
Neutral	Little (Sad, Fearful, Exciting, Relax) Ecstasy, Mellow	599
Relaxation	Calm, Chill, Relaxing	574
Sad	Hate, Depressing, Melancholic, Sentimental, Shameful, Distress, Anguish	498
Tension	Anger, Hate, Rage	528

Fig. 1 Proposed music video dataset samples. From first row, three samples are represented for each emotion class *Exciting*, *Fear*, *Neutral*, *Relaxation*, *Sad* and *Tension* respectively

is generally created with a fast tempo, high loudness, and irregular rhythm. *Relaxation* is a state of low tension or return in equilibrium state. The visual information in this category generally includes the natural scenario and musical instruments. The music in this category generally have slow tempo and complementing harmonies. *Sad* is an unpleasant feeling of human psychology that can be a result of mental suffering or hurt a human being. This state of emotion can be categorized in visual appearance by characterizing the feeling of loss, despair, grief, sorrow or disappointment. The sad music generally has a slow tempo and minor tonality. The *Tension* category includes violent scenes that are most likely to be highly arousing and elicit negative emotions. The music with tension emotion generally includes high loudness, fast tempo, and clashing harmonies. The *Neutral* category of human emotion may have influence of other basic emotion classes so the visual and acoustic information can have a mixed representation of emotions according to time and frame feature. The music video dataset[1] and the related experiments[2] are publicly available on GitHub (https://github.com/) to attract new researchers in this area.

3 Method

This section covers preprocessing for network input, transfer learning, unimodal and multimodal approach for music and video, and late fusion for emotion classification of deep neural network.

3.1 Preprocessing for network input

The music data preprocessing for 1D convolutional neural networks (CNN) first need zero padding to make the full-length audio waveform. The 30-s music signal is sampled at 16 kHz which corresponds to a 480,000-dimensional input vector to the 1D CNN music network as shown in Fig. 3. The pre-trained 2D music CNN [6] also need zero padding to make the full-length audio that can generate the fixed size Mel spectrogram (96-mel bins × 1876 temporal frames). A Mel spectrogram is a 2D representation of frequency content over time found by

[1] https://github.com/yagyapandeya/Music_Video_Emotion_Dataset
[2] https://github.com/yagyapandeya/Music_Video_Emotion_Recognition

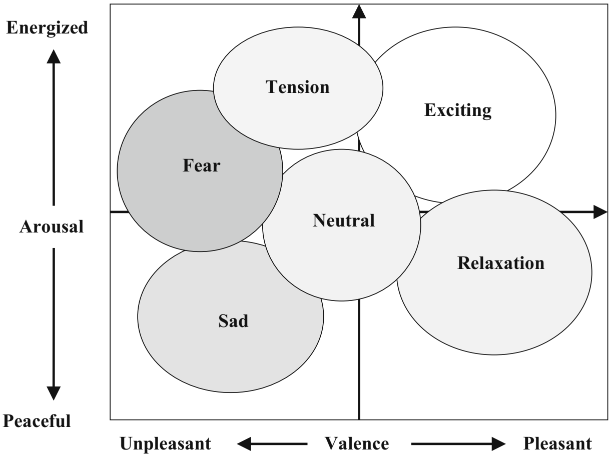
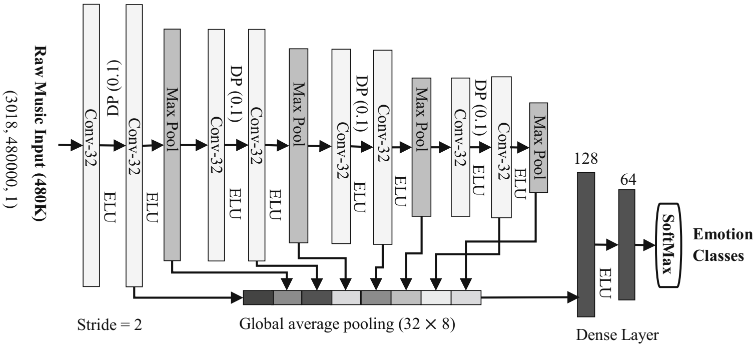


Fig. 2 A Valence-arousal emotion space for music-video content analysis

taking the absolute value of the short-time Fourier Transform (STFT). The other hyperparameters are kept the same as in [6] for input processing and network fine-tuning. Two video networks, namely, 3D convolutional network (C3D) and the Inflated Inception-V1 (I3D) are used in this research for visual emotion classification. Before input to network, all video frames should have adjusted into proper size, number and channel. The C3D and I3D networks are trained using 32 frames of video with red, green and blue channels. All the video frames are extracted in uniform time intervals to capture whole video information contents. The hyperparameter setting, temporal convolution size, and size of max pooling across the channel are kept the same as defined in respective pre-trained video networks.

3.2 Audio architectures

The effectiveness of CNNs on a variety of tasks lies in their capability to learn features from raw data in an end-to-end pipeline for targeting a particular task. Many recent music processing networks use the magnitude representations of Mel spectrogram of music and neglect the phase information. The way to preserve both magnitude and phase information of the auditory



signal is to use raw waveform as input to audio networks. Several audio architectures with raw waveform input are proposed for music classification in past [9, 31], but there is still no clear front running architecture for music emotion recognition. In this experiment, we proposed an audio architecture, shown in Fig. 3, for music emotion analysis with raw waveform input. The key component of our model is the 1D convolution operation with small-sized filters (3 × 1) over a long vector representation of music. We perform 32 convolutions with stride 2 and max-pooling (MP) across the channels with a pooling size of 2. This reduces the dimensionality of the signal while preserving the necessary statistics in the convolved signal. The exponential linear unit (ELU) [7] is used as an activation function. The dropout (DP) [50] with a probability of 0.1 is adopted in the training for alleviating overfitting. We collect the diverse network features using global average pooling (GAP) from various network layers and pass them to fully connected layers. Although the proposed music emotion network has comparatively lower evaluation results due to lack of training data but it plays a supportive role in a multimodal environment.

Another music network used in this research is a pre-trained audio network [6] with Mel spectrogram input. Although spectrogram representation of music ignores the audio phase information, the networks trained with large audio samples obviously have the generalization capability, and hence provides a facility of transfer learning [6, 43]. We use GAP from each convolutional layer and concatenate all the vectors and then fine-tune the whole network with our music data for emotion classification. The ReLUs (Rectified Linear Units) [19] is utilized to replace the ELU units and Adam optimizer with a learning rate of 0.001 applied for fine-tuning the network. The audio CNN block in Figs. 4 and 5 shows the 2D music emotion detection network.

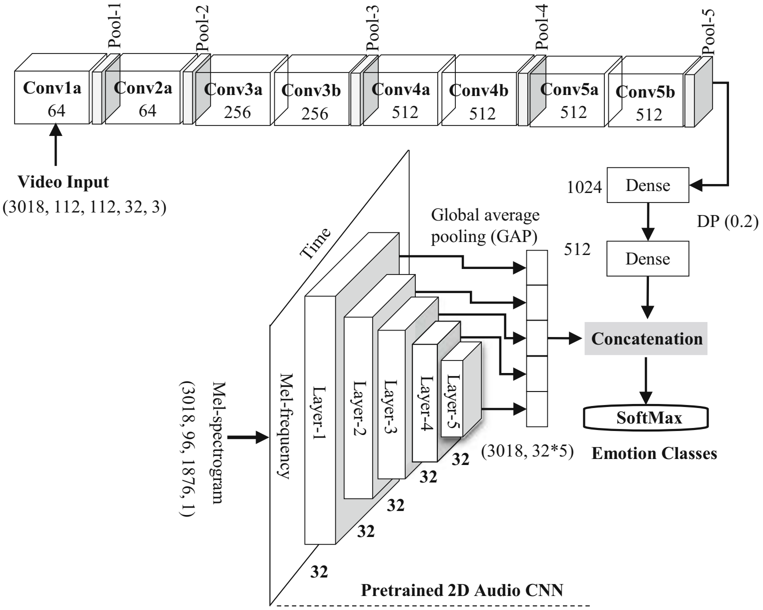


Fig. 4 Best performing music video multimodal architecture: C3D with pre-trained audio CNN with spectrogram input

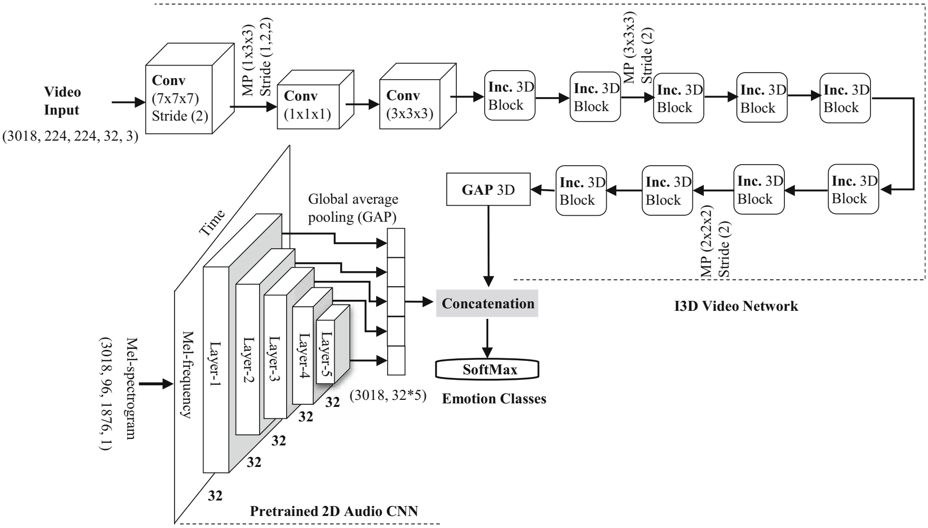


Fig. 5 Music video Multimodal architecture: I3D with pre-trained audio CNN with spectrogram input

3.3 Video architectures

Many works focusing on applying CNN models to the video domain with an aim to learn hidden spatiotemporal patterns. In video classification, researchers come up with methods in three-dimensional convolutional neural networks (3DCNN). These networks operate on stacked video frames and extend the original 2D convolutional kernel into 3D kernel to capture both spatial and temporal information. Although the 3D video models are complex and time-consuming during training, they have achieved the best result in the video analysis. Therefore, C3D and I3D video architecture are borrowed for video emotion analysis in this research.

The last two fully connected layers of the original C3D network are modified with lower dimension and DP with probability 0.2 is applied for alleviating overfitting in fine-tuning with our music video data. Stochastic gradient descent (SGD) [3] optimizer with learning rate 0.00001 and no momentum performs the best in the C3D network during finetuning for emotion categorization. The Adam [28] optimizer with learning rate 0.00001 also performs well for C3D but cannot exceed the SGD on training data and multimodal feature fusion. The modified structure of the C3D network for video emotion classification is illustrated in Fig. 4 (video network only). In the case of the I3D network, the output of the final inception block is passed through 3D global average pooling, and the whole network is fine-tuned with our music video data. For the I3D network, fine-tuning with Adam optimizer with learning rate 0.0001 perform the best and other hyperparameters are kept the same as defined in the original paper [4]. The detail of I3D architecture is shown in the upper right part of Fig. 5 (only the video processing network). Both C3D and I3D video networks are also fine-tuned using RMSprop [20] optimizer with learning rate 0.001 but it shows worse performance than the other optimizer.

3.4 Multimodal emotion classification

Recently, the architectures for multimodal machine learning made attraction that mostly integrates multimedia data publicly available on the Internet. Any multiple sources can be merged for a multimodal approach, but we choose here only the audio and video information. We fuse the decision-level features of respective C3D and I3D video networks with those of 1D Music CNN and 2D Music CNN, which results out the four multimodal architectures namely, *C3D plus 1D Music CNN*, *I3D plus 1D Music CNN*, *C3D plus 2D Music CNN*, and *I3D plus 2D Music CNN*. Each unimodal emotion classifier for audio and video is first fine-tuned separately with our emotion dataset. The final SoftMax classifier of each unimodal is removed and then we use it again as a multimodal feature classifier for final multimodal feature fusion. To decrease variance, and bias or improve predictions, the decision level feature of all proposed multimodal are combined into one predictive model, named the modal as *Integrated multimodal*. The parallel decision of integrated multimodal outperforms all other proposed multimodal.

3.5 Transfer learning

Transfer learning is a transformation process of any learned knowledge from one or more source domains to a target domain so that the target can enhance its prediction capability to solve the similar problems [53, 57]. Transfer learning is a proven technology in visual information processing [17, 22, 51] that helps to boost the system performance when there is only a small number of labeled training dataset. The studies in [6, 12, 30, 42, 43, 63] show versatile use of transfer learning in the diversified acoustic environment. To overcome the lack of data problem we also adopt the transfer learning from some well-known deep neural networks for music and video classification. We first load the pre-trained weights and then fine-tune the source neural network to make it be adapted to our music video dataset. After then, the learned features of each unimodal emotion classification are extracted for the final multimodal decision. In this research two pre-trained video networks, namely, C3D trained on sport-1 M dataset [25] and I3D trained on RGB ImageNet [10] and kinetic dataset [26], are used. The C3D and I3D networks are fine-tuned with subsampled 32 frames that our GPU server can cope with. A pre-trained 2D Music CNN [14] trained with a Million song dataset [35] is adopted as a music emotion classifier. We also fine-tune this audio network to generalize the network capability for music emotion classification. A little variation is made on original pre-trained networks for music video emotion classification, which is described in Section 3.3 and 3.4.

3.6 Decision-level feature fusion

Multimodal fusion is the process of integrating information from multiple sources for classification or regression tasks. There have been three information fusion methods including early, late and hybrid fusion. As in [11, 41, 69], the multimodal fusion provides the benefits of robustness, complementary information gain and functional continuity of system even in the failure of one or more modalities. Early (or feature-level) fusion integrates low-level features from each modality by correlation, which potentially accomplishes better task, but has difficulty in temporal synchronization among various input sources. The late (or decision-level) fusion obtains unimodal decision values and integrates them to obtain the final decision.

Although the late fusion ignores some low-level interactions between modality, it allows easy training with more flexibility, and simplicity to make predictions when one or more of modalities are missing. The hybrid (or mid-level) fusion attempts to exploits the advantages of both early and late fusion in a common framework. In this research, we exploit the late fusion, in which the highest level pre-trained features are combined to make a final decision by a SoftMax layer. The reason why we cannot use the early fusion is that the data scarcity made us use the transfer learning method where it should keep consistent the original structure of pre-trained audio and video network. Note that each of the pre-trained networks does not allow any low-level information fusion facility because each neural network is designed for its particular task so that the input and the low-level structure cannot be modified during transfer learning. Hence, we concatenate the learned features of each unimodal network for separate music and video emotion decisions and make a final decision by a SoftMax layer.

4 Result and discussion

In this research, we select some front-line audio and video models to test our proposed music video emotion dataset. We use the transfer learning technique and fine-tune on pre-trained CNNs to adopt them with our proposed music video dataset. We use three optimization techniques in various CNN architectures and set their various parameters for music video emotion prediction. Table 2 illustrates the evaluation results of our various unimodal classification networks and optimizer's influence over various learning factors. It is hard to make a concrete decision of the best optimizer in terms of fixed hyperparameter value and some evaluation matric. There is no certain rule to select the best parameter values for an optimizer, but we found some parameter values (illustrates in Table 2) that perform relatively better in this experiment.

The results of audio and video unimodal are integrated for the final multimodal structure. In the case of the C3D video structure, the Adam and SGD optimizer give a

Table 2 The effect of optimizer in unimodal classifier of music or video network

Optimizer	1D Music CNN	2D Music CNN	C3D Video Network	I3D Video Network
Adam	LR: *0.001* Test Accuracy: **0.4453** F1-score: **0.41** ROC AUC Score: **0.757**	LR: *0.001* Test Accuracy: 0.6539 F1-score: 0.65 ROC AUC Score: 0.916	LR: *0.00001* Test Accuracy: 0.6423 F1-score: **0.64** ROC AUC Score: **0.898**	LR: *0.0001* Test Accuracy: **0.6622** F1-score: **0.66** ROC AUC Score: **0.899**
SGD	LR: *0.001* M: *0.5* Test Accuracy: 0.4238 F1-score: 0.37 ROC AUC Score: 0.734	LR: *0.001* M: *0.5* Test Accuracy: **0.7251** F1-score: **0.72** ROC AUC Score: **0.934**	LR: *0.00001* M: *0.0* Test Accuracy: **0.6440** F1-score: **0.64** ROC AUC Score: 0.890	LR: *0.001* M: *0.5* Test Accuracy: 0.6026 F1-score: 0.59 ROC AUC Score: 0.879
RMSprop	LR: *0.001* Test Accuracy: 0.4304 F1-score: 0.39 ROC AUC Score: 0.748	LR: *0.001* Test Accuracy: 0.6887 F1-score: 0.68 ROC AUC Score: 0.921	LR: *0.001* Test Accuracy: 0.5678 F1-score: 0.56 ROC AUC Score: 0.848	LR: *0.001* Test Accuracy: 0.5281 F1-score: 0.53 ROC AUC Score: 0.832

The bold number represent the highest evaluation score

LR, learning rate *and M*, momentum

very similar result, but we select the SGD result for multimodal integration. In the case of audio structure, the 2D Music CNN performs better than the 1D Music CNN because the network is pre-trained on Million Song Dataset. Even the 1D Music CNN includes the phase and magnitude to audio stream, the network cannot outperform the 2D Music CNN because the data samples are very limited for end-to-end training. In video network architecture, the I3D perform better than C3D network because of deep architecture and more generalized feature.

The performance evaluation metrics used in our experiment are accuracy, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). The accuracy refers to the percentage of correctly classified unknown data samples, and F1-score computes harmonic mean between precision and recall. ROC (receiver operating characteristic) curve is a graph showing the performance of a classification model at all classification thresholds with true and false-positive rates. The ROC-AUC score measures the entire two-dimensional area underneath the ROC curve, which provides an aggregated measure of performance across all possible classification thresholds.

We integrate the learned features from our best performing unimodal classifiers over various optimizers for music video emotion prediction. The music networks are integrated with video networks at the decision level and classify the concatenated features using the SoftMax emotion classifier. The classification is done by six-fold cross-validation. Table 3 illustrates the results of various multimodal combinations with learned features obtained from respective audio and video unimodal classifiers. The result of decision-level feature fusion shows that the best result is obtained when all audio and video features are combined with a SoftMax decision operator. The confusion matrix from six-fold cross-validation is shown in Fig. 6 that visualize the performance of each multimodal classifier. The confusion matrix of our various combinations for multimodal classifiers includes the number of samples confused to each other. All the confusion matrixes of Fig.6 illustrate that the ‘*Relaxation*’ and ‘*Sad*’ emotion class are more confusing to each other because they all are expressed in similar calm state. The ‘*Exciting*’ and ‘*Tension*’ emotion classes are well classified compared to other silent emotion classes because the body action captured by the video network plays a vital role for it. As the ‘*Neutral*’ class include the common characteristic of all emotion categories, it seems more confused by all proposed model. To better understand our system performance, we visualize the ROC curve, shown in Fig. 7, for all multimodal classifiers and their integrated structure. It illustrates that the integrated multimodal fuse all learned features of proposed multimodal in decision level and hence produces the best ROC-AUC score.

Table 3 Best operating multimodal for music video emotion prediction

Metrics		C3D + 1D Music CNN	I3D + 1D Music CNN	C3D + 2D Music CNN	I3D + 2D Music CNN	Integrated multimodal
Test Set Accuracy (percentage)	Minimum	68.389	66.003	81.709	83.300	**86.282**
	Mean	70.941	69.781	84.956	84.426	**88.568**
	Maximum	72.962	72.166	87.872	85.884	**89.860**
F-score		0.70	0.69	0.84	0.84	**0.88**
ROC AUC Score		0.917	0.925	0.979	0.977	**0.987**

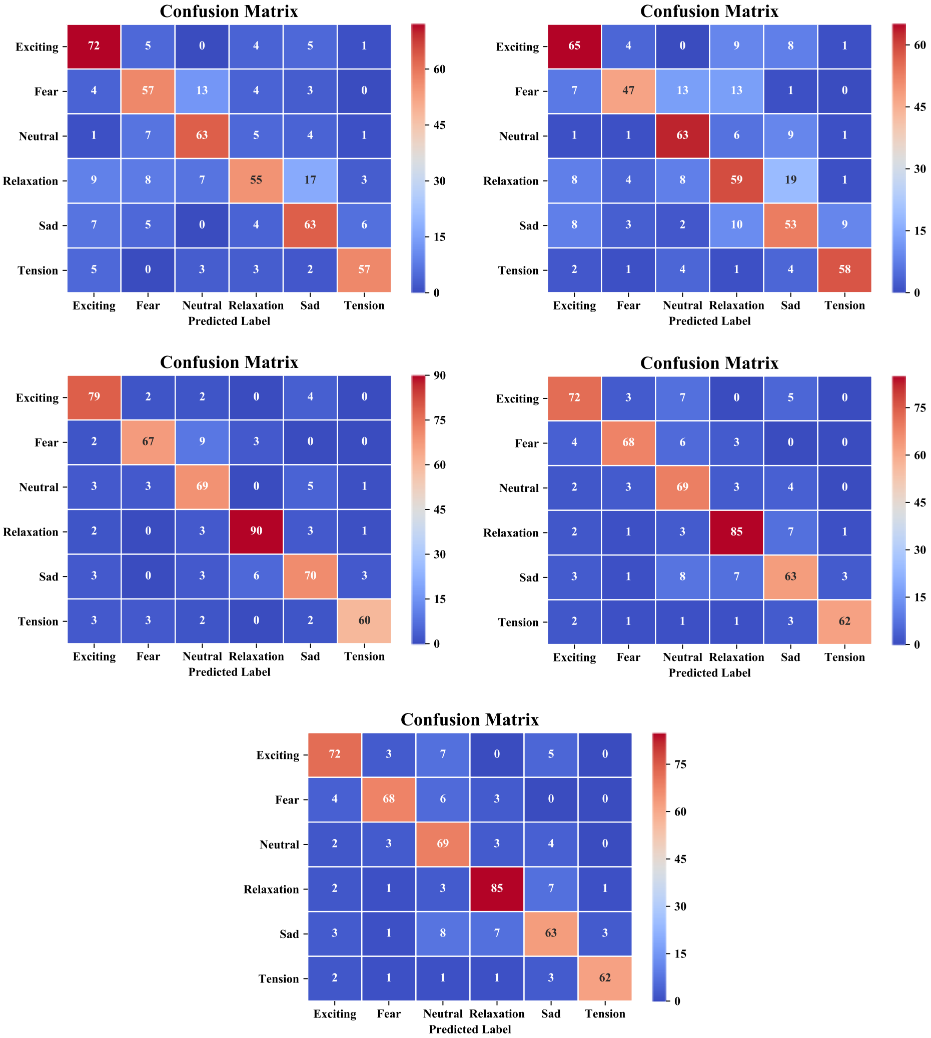


Fig. 6 Confusion matrix of multimodal: C3D + 1D Music CNN (top left), I3D + 1D Music CNN (top right), C3D + 2D Music CNN (2nd row bottom left) and I3D + 2D Music CNN (2nd row bottom right), Integrated multimodal (bottom)

The evaluation results show that any multimodal (music plus video) classifier boosts the overall performance, and reduce the rate of confusion, as we expected. In the case of the 1D music CNN network, even the features classification accuracy is comparatively low because of the lack of training data, but it is supportive for any video unimodal classifier to enhance overall performances. The pre-trained networks are trained on huge data samples, so they made better performance on the emotion classification of music video dataset.

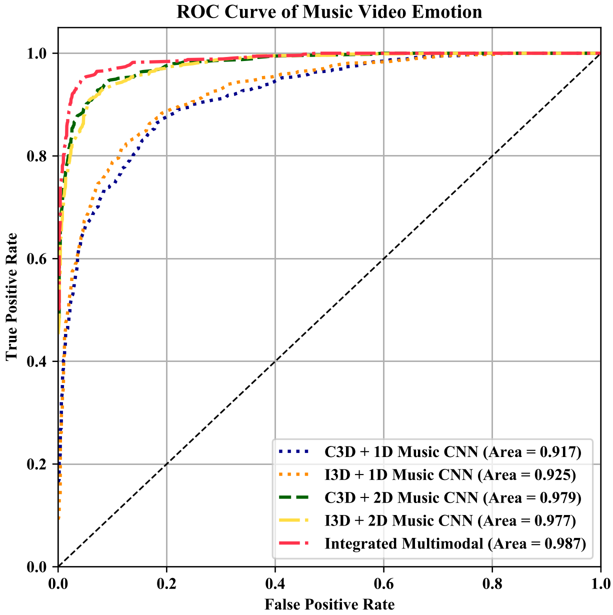


Fig. 7 ROC curve of multimodal used in this research. Each curve is represented by a unique color and the corresponding AUC score is given in parentheses

5 Summary

The music-video contents are a rich source of human emotion. Music video emotions are central to how we perceive the visual and acoustic environment concurrently, how we make sense of it, and how we make timely decisions. The visual sentiments (face expression, body action, eye states) and music structural features (tempo, mode, melody, rhythm, lyric, and loudness) are jointly represented on a music video that makes its study complex and nuanced. Even though humans are known to perceive hundreds of different emotions, there is still little agreement on how best to categorize and represent emotions. The vague emotion boundary, personality and lack of labeled data are the reasons that make affective computing for music video challenging and hard to understand.

We show that the transfer learning and late decision-level fusion are useful to overcome the lack of data problem. Also, we effectively use the CNN-based structure to coarsely classify the nuanced emotions of music videos. Our study starts with the construction of a small music video emotion dataset. Then the music and video part are separated in order to use CNN structures pre-trained for other audio and video tasks. We try to fine-tune the audio and video networks separately or train 1D CNN for audio with the constructed small emotion dataset. This process for unimodal CNNs gives the proper features for multimodal emotion classification of music videos. Then the pooled features from each audio and video modality are finally combined by learned SoftMax operator to produce the coarsely grouped emotion category. This type of late fusion is inevitable to keep the pre-trained network structure consistent.

We made an analysis of various unimodal with various tuning parameters and the contribution of each CNN structure on integration with other structures for music video emotion classification. We evaluate five different multimodal networks using various metrics. The evaluation results show a boost in the performance of the multimodal architectures compared to each unimodal emotion classifier constructed in the fine-tuning stage. The decision based on features taken from all unimodal networks shows the best performance. The result suggests human high-level emotions are automatically well classified in the proposed CNN-based multimodal networks, even though a small amount of labeled data samples is available for training. Our different combinations of audio and video classifiers can be separately used to classify the emotion of a music video. In addition, it can be used as multi-agents to make an ensemble decision for music video emotion.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) under Basic Science Research Program NRF-2019R1A6A1A09031717 and Brain Korea 21 PLUS Project, National Research Foundation of Korea.

References

1. Bahuleyan H (2018) Music genre classification using machine learning techniques. arXiv:1804.01149v1
2. Baltrusaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning:a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41:423–443
3. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. Springer proceedings of COMPSTAT'2010 177–186
4. Carreira J, and Zisserman A (2018) Quo vadis, action recognition? A new model and the kinetics dataset. arXiv:1705.07750v3
5. Chang WY, Hsu SH, and Chien JH (2017) FATAUVA-net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. IEEE 2160-7516
6. Choi K, Fazekas G, Sandler M and Cho K (2017) Transfer learning for music classification and regression tasks. International Society for Music Information Retrieval Conference, Suzhou, China 141–149
7. Clevert DA, Unterthiner T and Hochreiter S (2016) Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289
8. Cowen AS, Keltner D (2017) Self-report captures 27 distinct categories of emotion bridged by continuous gradients. PNAS 114(38):E7900–E7909
9. Dai W, Dai C, Qu S, Li J, and Das S (2016) Very deep convolutional neural networks for raw waveforms. arXiv:1610.00087v1
10. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition:1063–6919
11. Ding W, Xu M, Huang D, Lin W, Dong M, Yu X, Li H (2016) Audio and face video emotion recognition in the wild using deep neural networks and small datasets. International conference on multimodal interfaces. Tokyo, Japan
12. Elshaer MEA, Wisdom S, Mishra T (2019) Transfer learning from sound representations for anger detection in speech. arXiv:1902.02120v1
13. Fan Y, Lu X, Li D, Liu Y (2016) Video-based emotion recognition using CNN-RNN and C3D hybrid networks. International conference on multimodal interfaces. Tokyo, Japan
14. Fridman L, Brown DE, Glazer M, Angell W, Dodd S, Jenik B, Terwilliger J, Patsekin A, Kindelsberger J, Ding L, Seaman S, Mehler A, Sipperley A, Pettinato A, Seppelt B, Angell L, Mehler B, and Reimer B (2019) MIT advanced vehicle technology study: large-scale naturalistic driving study of driver behavior and interaction with automation. arXiv:1711.06976v4

15. Gao Z, Xuan HZ, Zhang H, Wan S and Choo KKR (2018) Adaptive fusion and category-level dictionary learning model for multi-view human action recognition. IEEE Internet of Things Journal
16. Gao Z, Wang YL, Wan SH, Wang DY, Zhang H (2019) Cognitive-inspired class-statistic matching with triple-constrain for camera free 3D object retrieval. Futur Gener Comput Syst 94:641–653
17. Garces, MLE (2018) Transfer learning for illustration classification, arXiv:1806.02682v1
18. Grekow J (2018) From content-based music emotion recognition to emotion maps of musical pieces. Springer
19. Hahnloser RHR, Sarpeshkar R, Mahowald MA, Douglas RJ, and Seung SH (2000) Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405-6789-947
20. Hinton G, Srivastava N, and Swersky K (2012) Lecture 6d - a separate, adaptive learning rate for each connection. Slides of Lecture Neural Networks for Machine Learning.
21. Hong S, Im W, and Yang HS (2017) Content-based video–music retrieval using soft intra-modal structure constraint. arXiv:1704.06761v2.
22. Hussain M, Bird JJ, Faria DR (2018) A study on CNN transfer learning for image classification. UKCI 2018: Advances In Intelligent Systems and Computing, (840) 191-202 Springer
23. Kahou SE, Bouthillier X, Lamblin P, Gulcehre C and at al. (2015) EmoNets: Multimodal deep learning approaches for emotion recognition in video. arXiv:1503.01800v2.
24. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. IEEE conference on Computer Vision and Pattern Recognition:1725–1732
25. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition
26. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, and Zisserman A (2017) The kinetics human action video dataset. arXiv:1705.06950
27. Kaya H, Gürpınar F, Salah AA (2017) Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis Comput 65:66–75
28. Kingma D and Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
29. Koelstra S, M¨uhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt N, and Patras I (2012) DEAP: a database for emotion analysis using physiological signals. IEEE Trans Affect Comput
30. Kunze J, Kirsch L, Kurenkov I, Krug A, Johannsmeier J, and Stober S (2017) Transfer learning for speech recognition on a budget. arXiv:1706.00290v1
31. Lee J, Park J, Kim KL, Nam J (2018) SampleCNN: end-to-end deep convolutional neural networks using very small filters for music classification. Applied science. https://doi.org/10.3390/app8010150
32. Liu X, Chen Q, Wu X, Yan L, Ann Yang L (2017) CNN based music emotion classification. arXiv: 1704.05665
33. Lövheim H (2012) A new three-dimensional model for emotions and monoamine neurotransmitters. Med Hypotheses 78:341–348
34. Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio-visual emotion fusion (AVEF): a deep efficient weighted approach. Information Fusion 46:184–192
35. Mahieux TB, Ellis DP, Whitman B, and Lamere P (2011) The million song dataset. 12th international conference on music information retrieval, Miami FL 591-596
36. Minaee S and Abdolrashidi A (2019) Deep-emotion: facial expression recognition using attentional convolutional network. arXiv:1902.01019v1
37. Ng JY, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. IEEE conference on computer vision and pattern recognition:4694–4702
38. Nguyen D, Nguyen K, Sridharan S, Ghasemi A, Dean D and Fookes C (2017) Deep spatio-temporal features for multimodal emotion recognition. IEEE Winter Conference on Applications of Computer Vision
39. Noroozi F, Sapiński T, Kamińska D, Anbarjafari G (2017) Vocal-based emotion recognition using random forests and decision tree. International Jornal of Speech Technology 20:239–246
40. Ortega JDS, Senoussaoui M, Granger E, and Pedersoli M (2019) Multimodal fusion with deep neural networks for audio-video emotion recognition. arXiv:1907.03196v1.
41. Ouyang X, Kawaai S, Goh EGH, Shen S, Ding W, Ming H, Huang DY (2017) Audio-visual emotion recognition using deep transfer learning and multiple temporal models. International conference on multimodal interfaces. Glasgow, UK
42. Pandeya YR, Lee J (2018) Domestic cat sound classification using transfer learning. International Journal of Fuzzy Logic and Intelligent Systems 18-2:154–160
43. Pandeya YR, Kim D, and Lee J (2018) Domestic cat sound classification using learned features from deep neural nets. Applied science 1949

44. Pini S, Ben-Ahmed O, Cornia M, Baraldi L, Cucchiara R, Huet B (2017) Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. International conference on multimodal interfaces. Glasgow, UK
45. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Information Fusion 37:98–125
46. Ringeval F, Sonderegger A, Sauer J, and Lalanne D (2013) Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).
47. Rozgic V, Vitaladevuni SN, Prasad R (2013) Robust EEG emotion classification using segment level decision fusion. IEEE International Conference on Acoustics, Speech and Signal Processing
48. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39-6:1161–1178
49. Shiqing Z, Shiliang Z, Huang T, Gao W, Tian Q (2018) Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology:28–10
50. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15-1:1929–1958
51. Su YC, Chiu TH, Yeh CY, Huang HF, and Hsu WH (2015) Transfer Learning for Video Recognition with Scarce Training Data for Deep Convolutional Neural Network. arXiv:1409.4127v2
52. Sun K, Yu J, Huang Y, and Hu X (2009) An improved valence-arousal emotion space for video affective content representation and recognition. IEEE International Conference on Multimedia and Expo
53. Tan C, Sun F, Kong T, Zhang W, Yang C, and Liu C (2018) A survey on deep transfer learning. arXiv: 1808.01974v1
54. Thayer RE (1989) The biopsychology of mood and arousal. Oxford University Press
55. Tian H, Tao Y, Pouyanfar S, Chen SC, Shyu ML (2019) Multimodal deep representation learning for video classification. World Wide Web 22:1325–1341
56. Tiwari SN, Duong NQK, Lefebvre F, Demarty CH, Huet B and Chevallier L (2016) Deep features for multimodal emotion classification. HAL-01289191.
57. Torrey L, Shavlik J (2009) Transfer learning. IGI Global Publication Handbook of Research on Machine Learning Applications
58. Tran D, Bourdev L, Fergus R, Torresani L, and Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. IEEE International Conference on Computer Vision 4489–4497
59. Tremblay J, To T, Sundaralingam B, Xiang Y, Fox D, and Birchfield S (2018) Deep object pose estimation for semantic robotic grasping of household objects. arXiv:1809.10790v1
60. Tripathi S, Acharya S, and Sharma RD (2017) Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. Twenty-Ninth Association for the Advancement of Artificial Intelligence Conference on Innovative Applications
61. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, and Zafeiriou S (2017) End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing 1301-1309
62. Wang S, Ji Q (2015) Video affective content analysis: a survey of state-of-the-art methods. IEEE Trans Affect Comput
63. Wang D, Zheng TF (2015) Transfer learning for speech and language processing. APSIPA Annual Summit and Conference 2015
64. Wu H, Chen Y, Wang N, and Zhang Z (2019) Sequence level semantics aggregation for video object detection. arXiv:1907.06390v2
65. Xu YS, Fu TJ, Yang HK, Lee CY (2018) Dynamic video segmentation network. arXiv:1804.00931v2
66. Yang YH and Chen HH (2012) Machine recognition of music emotion: a review. ACM transactions on intelligent systems and technology 3-3-40
67. Zhang L and Zhang J (2018) Synchronous prediction of arousal and valence using LSTM network for affective video content analysis. arXiv:1806.00257

68. Zhang L, Tjondronegoro D, Chandran V (2014) Representation of facial expression categories in continuous arousal–valence space: feature and correlation. Image Vis Comput 32:1067–1079
69. Zhang S, Zhang S, Huang T, Gao W (2016) Multimodal deep convolutional neural network for audio-visual emotion recognition. ACM on international conference on multimedia retrieval 281-284.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yagya Raj Pandeya was born in Banlek village, Dadeldhura, Nepal in 1988. He receives the B.E. and M.E. degree in Computer Engineering from the Pokhara University of Nepal, in 2010 and 2013, respectively. He was Head of the Department of Computer Engineering at NAST College in Dhangadhi, Nepal. He joins the Ministry of Home Affairs Nepal from 2015 to 2017. Mr. Yagya is currently a Ph.D. fellow at Fuzzy Logic and Artificial Intelligence Laboratory at Chonbuk National University, Korea. His research interests include audio-video information retrieval, audio event detection and localization, and animal sound behavior analysis.

Joonwhoan Lee received his BS degree in Electronic Engineering from the University of Hanyang, Korea in 1980. He received his MS degree in Electrical and Electronics Engineering from KAIST, Korea in 1982, and the Ph.D. degree in Electrical and Computer Engineering from the University of Missouri, the USA in 1990. He is currently a Professor in the Department of Computer Engineering, Jeonbuk National University, Korea. His research interests include image and audio processing, computer vision, emotion engineering, etc.

Affiliations

Yagya Raj Pandeya[1] • Joonwhoan Lee[1]

Yagya Raj Pandeya
yagyapandeya@gmail.com

[1] Division of Computer Science and Engineering, Jeonbuk National University, Jeonju, South Korea