sensors

MDPI

Article

Deep-Learning-Based Multimodal Emotion Classification for Music Videos

Yagya Raj Pandeya, Bhuwan Bhattarai and Joonwhoan Lee *

Department of Computer Science and Engineering, Jeonbuk National University, Jeonju-City 54896, Korea; yagyapandeya@gmail.com (Y.R.P.); bhubon240@gmail.com (B.B.)
* Correspondence: chlee@chonbuk.ac.kr

Abstract: Music videos contain a great deal of visual and acoustic information. Each information source within a music video influences the emotions conveyed through the audio and video, suggesting that only a multimodal approach is capable of achieving efficient affective computing. This paper presents an affective computing system that relies on music, video, and facial expression cues, making it useful for emotional analysis. We applied the audio–video information exchange and boosting methods to regularize the training process and reduced the computational costs by using a separable convolution strategy. In sum, our empirical findings are as follows: (1) Multimodal representations efficiently capture all acoustic and visual emotional clues included in each music video, (2) the computational cost of each neural network is significantly reduced by factorizing the standard 2D/3D convolution into separate channels and spatiotemporal interactions, and (3) information-sharing methods incorporated into multimodal representations are helpful in guiding individual information flow and boosting overall performance. We tested our findings across several unimodal and multimodal networks against various evaluation metrics and visual analyzers. Our best classifier attained 74% accuracy, an f1-score of 0.73, and an area under the curve score of 0.926.

Keywords: channel and filter separable convolution; end-to-end emotion classification; unimodal and multimodal

Citation: Pandeya, Y.R.; Bhattarai, B.; Lee, J. Deep-Learning-Based Multimodal Emotion Classification for Music Videos. *Sensors* **2021**, *21*, 4927. https://doi.org/10.3390/s21144927

Academic Editors: Soo-Hyung Kim and Gueesang Lee

Received: 14 June 2021
Accepted: 17 July 2021
Published: 20 July 2021

1. Introduction

Emotion is a psycho-physiological response triggered by the conscious or unconscious perception of external stimuli. There is a wide variety of factors associated with emotion, including mood, physical feeling, personality, motivation, and overall quality of life. Emotions play a vital role in decision making, communication, action, and a host of cognitive processes [1]. Music videos are commercial products that pair music with imagery to promote a song or album. Music videos convey affective states through verbal, visual, and acoustic cues. Because they blend multiple types of information, a number of different methods of analysis are needed to understand their contents. In the context of music videos, identifying emotional cues requires not only analysis of sound, but visual information as well, including facial expressions, gestures, and physical reactions to environmental changes (e.g., changes in color scheme, lighting, motion, and camera focus points).

A number of studies have attempted to show how music carries human affective states [2] and boosts mood and self-confidence [3–5]. Sometimes, this emotional effect is counterintuitive, as even sad music can evoke pleasure and comfort in listeners [6,7]. Pannese et al. [8] used the conceptual metaphor of a chain of musical emotion that emanates from a recording or performance and works its way to the audience and listeners. The performer or composer transmits emotion at the production level. The music, then, evokes emotion at the level of perception when it is received by the audience or listeners. Finally, an affective state is brought about in the audience or listener in response to the song at the induction level. These authors conceived of emotion using a top-down approach, and were unable to describe psychological changes in affective states when listening to music.

The fact that the act of listening and responding to music involves subjective assessments on the part of the listener adds to the complexity and uniqueness of affective computing. Several musical components, including harmony, tonality, rhythm, mode, timing, melody, loudness, vibrato, timbre, pitch, and the vocalist's performance, make each musical work unique. Visual components further complicate emotional analysis, as performers can add emotions to their music through body movements, gestures, and facial expressions.

With the rise of the Internet and social media, users are engaging more than before with multimedia content in order to communicate and explore their affective states. In this paper, we focused on analyzing music and image sequences produced by either the music industry or random social media users. Music videos are among the most circulated types of content on the Internet. Emotion recognition has already been employed by music streaming services, the video game industry, the advertising industry, the television industry, and the music industry for audio–video content matching and synchronization [9]. Emotion recognition is also used in mental treatment, meditation, and encouragement training [10].

Deep learning technology makes our daily lives more convenient across diversified domains. Numerous success stories have rapidly spread in the domains of animal sounds [11–14] and music information retrieval [15]. Automatic affective computing of music using deep neural networks (DNNs) allows for the processing of a large number of features from multiple modalities. However, DNNs require a large amount of data for training, and data scarcity has held back research in this field. Other major challenges for emotional analysis in music videos include the undefined frameworks for emotional representation, a lack of standard datasets, and difficulties in annotation due to subjective and affective states that vary across time. Moreover, how individuals demonstrate their emotions varies across cultural groups, languages, and music makers. To compound the problem, multiple information sources (audio, video, facial expressions, and lyrics) are used to communicate information about affective states. Finally, user-generated music videos may not present a consistent picture of emotion across audio and video sources. Annotators, in turn, must consider correlated sources to provide precise annotations.

This article seeks to enhance and improve a supervised music video dataset [16]. The dataset includes diversified music video samples in six emotional categories and is used in various unimodal and multimodal architectures to analyze music, video, and facial expressions. The unimodal term in this paper is used for a network that uses only one source of information, such as music, video, or facial expressions. The integrated structure of more than one unimodal source is termed as multimodal. We conducted an ablation study on unimodal and multimodal architectures from scratch by using a variety of convolution filters. The major contributions of this study are listed below:

(a) We extended and improved an existing music video dataset [16] and provided emotional annotation by using multiple annotators of diversified cultures. A detailed description of the dataset and statistical information is provided in Section 3.
(b) We trained unimodal and multimodal architectures with music, video, and facial expressions using our supervised data. The networks were designed using 2D and 3D convolution filters. Later, the network complexity was reduced by using a novel channel and separable filter convolution.
(c) An ablation study was conducted to find a robust and optimal solution for emotion classification in music videos. Music was found to be the dominant source for emotion-based content, and video and facial expressions were positioned in a secondary role.
(d) The slow–fast network strategy [17] was applied for multimodal emotion classification. The slow and fast branches were designed to capture spatiotemporal information from music, video, and facial expression inputs. The learned features of two parallel branches of a slow–fast network were shared and boosted by using a multimodal transfer module (MMTM) [18], which is an extension of "squeeze and excitation" (SE) [19].

We overcame the difficulties of input processing for multimodal data of large dimensions. Different networks were trained and interpreted by analyzing different information sources individually and jointly. The network performance was evaluated visually and statistically by using different evaluators.

The remainder of this article is organized as follows. In Section 2, we present related past studies on music, video, facial expression, and multimodal processing for affective computing. Section 3 explains our music video emotion dataset and the proposed deep neural network architectures. Section 4 includes the statistical and visual test results for the unimodal and multimodal architectures. Finally, Section 5 concludes the study by discussing further directions for future research.

2. Related Works

Different approaches to the analysis of emotions in music have been taken in the past by using data-driven explorations. At the time of writing, the authors were not aware of any research on deep learning that leveraged affective computing of music videos from scratch. Some unimodal architectures were studied well by using diverse datasets in the past. In this section, we discuss some existing techniques for affective computing of music, video, and facial expressions.

Music emotion recognition (MER) is one research sub-domain of music information retrieval (MIR) that focuses on the study of the characteristics of music and their correlation with human thinking. The audio can be dominant information for high-level semantics, such as emotion, in multimedia content. Lopes et al. [20] showed the importance of sound in horror movies or video games for the audience's perception of a tense atmosphere. Particularly for emotions in music, some research has shown satisfactory results by using music audio and lyrics. Naoki et al. [21] integrated their analysis of music lyrics and audio by using a mood trajectory estimation method. Algorithms used to generate these features from audio and to classify them include k-nearest neighbors, random forests, support vector machine (SVM), among other regression models. Song et al. [22] proposed a LastFM dataset classified into four emotional categories (angry, happy, sad, and relaxed) and used an SVM with polynomial and radial basis function kernels for classification. Other studies [23–25] used various handcrafted music features for affective computing. The comprehensive work of [26] proposed a dataset and compared various machine learning and deep learning methods. Recurrent neural networks [27,28] and convolutional neural networks (CNNs) [29] generally rely on time–frequency spectral representation for emotion classification in music. Tsunoo et al. [30] used rhythm and bass-line patterns to classify the music contained in the Computer Audition Lab 500-song (CAL500) dataset [31] into five emotional categories. The CAL500 dataset includes 500 popular songs from Western countries with semantic labels derived from human listeners. One CNN-based music emotion classification [32] method for the CAL500 dataset, as well as its enriched version (CAL500exp) [33], was used for the classification of 18 emotion tags in the dataset. Recently, after music source separation [34] and attention [35], individual music sources were also applied to improve prediction of emotions in music by using a spectral representation of audio as the input. The spectrogram was a handcrafted magnitude-only representation without phase information. Orjesek et al. [36] addressed this problem by using a raw waveform input for their classifier. Our study used both the real (magnitude) and imaginary (phase angle) information from audio for emotion classification because several studies [37–39] have demonstrated that phase information improves the performance of both speech and music processing.

Video is an ordered sequence of images that include several visual clues for affective computing. While there is no deep-learning-related research on the visual representation of emotion in music videos, several studies [40–42] have examined emotion in user-generated videos. The main challenge in the analysis of these videos is the subjective nature of emotion and the sparsity of video frames in which emotion is expressed. By using attention mechanisms, some researchers [43,44] found that it is beneficial to boost the visual cues

for emotion in video. In sequential data, temporal information is important. Xu et al. [45] conducted a study on the capture of temporal or motion information by using an optical flow in parallel with the RGB channel. Spatiotemporal data processing is improved by using a slow–fast network [17], where the slow branch carries spatial information with fewer video frames, and the fast branch is responsible for capturing the temporal information with a large number of video frames. Multiple information processing paradigms have also been used on movie clips to capture diverse emotional cues with multimodal representation. Modern music videos use a wide range of filmmaking styles as marketing tools in order to promote the sale of music recordings. Many music videos present specific images and scenes from the song's lyrics, while others take a more thematic approach. To produce a music video, the director tries to create a visual representation that matches their subjective analysis of the emotion in the piece of music. The classification of emotions in movies by using audio and visual information was proposed in [46,47] with the use of low-level audio–video features. Affective computing of movies was later studied in [48] by using a neural network and face and speech data. Reinforcement learning was used in [49] for identifying funny scenes in movies. These techniques can be useful for affective computing of music videos and dealing with multimedia content.

Facial expressions can provide information about humans' feelings and thoughts, and they play a crucial role in interaction and decision making. Facial expressions can have universal qualities [50] and have potential applications in human–computer interaction, computer vision, and multimedia entertainment. The facial expressions of a music video actor are crucial for affective computing because he/she is guided by the video director to bring their body movement and expressions in line with the emotions expressed by the music. In relation to other visual cues, such as gestures, background scenery, color, and camera position, facial expressions provide clearer visual cues for emotional analysis [51,52]. Some deep-learning-based research [53–55] has achieved satisfying results by using facial expression for different applications. Facial expressions have been extensively studied in speech recognition and have been found to be beneficial for improving learning networks' capabilities [56–58]. Seanglidet et al. [59] proposed the use of facial expression analysis in music therapy; however, facial emotions have not been used in the study of emotions in music videos.

The wide proliferation of multimedia content that is posted online is increasingly pushing researchers away from conventional unimodal architectures and towards complex multimodal architectures. Some multimodal architectures [60,61] have been proposed for affective computing analysis of music videos by using machine learning technology. Pandeya and Lee [16] proposed a supervised music video emotion dataset for a data-driven algorithm and used late feature fusion of audio and video representations after transfer learning. We extended their dataset and incorporated additional emotional cues from music, video, and facial expressions. These information sources treat emotions individually with their own schema of emotion representation. Multimodal representation is one means of dealing with the complex problem of emotional analysis, where each information source within a music video influences the emotions conveyed through the acoustic and visual representation. Affective computing of music videos, however, has not been used to interpret multiple sources of information. Our model is unique inasmuch as it used music, video, and facial expression information. We present various architectures, convolution techniques, and applied information-sharing methods for emotional classification of music videos.

3. Data Preparations

A number of emotion representation frameworks have been proposed in the last decade. The categorical model [62] describes human emotions by using several discrete categories. Conversely, the dimensional model [63–65] defines emotions as numerical values over two or three psychological dimensions, such as valence, arousal, and dominance. Some variants of these frameworks have been applied in MER studies [34,35,66,67] on

music data. Several datasets have previously been proposed for supervised emotional analysis of music. Some of these datasets [68–70] follow the categorical model, providing several discrete categories of emotions, and other datasets [24,26,71] used the dimensional model to represent emotions as values in a 2D valence and arousal space. Similarly to those of music, some video emotion datasets [72,73] have been proposed that used the categorical model, and others [74,75] used the dimensional model.

The 2D valence–arousal framework is the most widely used framework for emotion representation in music. While this approach overcomes the problem of categorical limitations and ambiguities in search tags, the categories are vague, unreliable, and not mutually exclusive [76]. The categorical representation can be a better approach for an online streaming service system where the end-user usually makes their demands for their favorite music videos based on class categories, such as emotion tags, singers, or genres. We used the framework identified in [16], where cross-correlation among six basic emotions was explained for music videos, and we extended the dataset. This dataset originally had 720 Excitation, 519 Fear, 599 Neutral, 574 Relaxation, 498 Sad, and 528 Tension data samples. The training, testing, and validation sets were not defined separately, and multiple samples were taken from the same music video. This led to a problem of overfitting, as DNNs are clever in capturing the shortcuts. Shortcuts are decision rules that perform well on standard benchmarks, but fail to transfer to more challenging testing conditions, such as real-world scenarios [77]. In the music video, such shortcuts can be the outer frame of the video, channel logos, and opening or background music. Our updated dataset has nearly twice as much data, which was derived from a wider variety of samples. Most samples were not repeated from a single music video, and the training and testing samples were taken from distinct sources. The statistical layout of our dataset and the number of samples are presented in Table 1.

Table 1. Music video dataset with various adjectives and statistics in each emotion class.

Emotion Class	Emotion Adjectives	Training Samples	Validation Samples	Testing Samples
Excited	Happy, Fun, Love, Sexy, Joy, Pleasure, Exciting, Adorable, Cheerful, Surprising, Interest	843	102	50
Fear	Horror, Fear, Scary, Disgust, Terror	828	111	50
Neutral	Towards (Sad, Fearful, Exciting, Relax) Ecstasy, Mellow	678	99	50
Relaxation	Calm, Chill, Relaxing	1057	148	50
Sad	Hate, Depressing, Melancholic, Sentimental, Shameful, Distress, Anguish	730	111	50
Tension	Anger, Hate, Rage	652	84	50
Total		4788	655	300

We categorized the dataset into six distinct classes based on their corresponding emotional adjectives. The "Excited" class usually includes positive emotions. The visual elements of the "Excited" class include images of a smile on a face, movement of arms, dancing scenes, bright lighting, and coloring effects. The audio components of this class include high pitch, large pitch variations, uniform harmony, high volume, low spectral repetition, and diverse articulations, ornamentation, and vibrato. The visual features of the "Fear" class reflect negative emotions via a dark background, unusual appearance, wide eyes, open mouth, a visible pulse on the neck, elbows directed inward, and crossed arms. Common visual elements in the "Tension" class are fast-changing visual scenes, crowded scenes, people facing each other, aggressive facial expressions with large eyes and open mouths, and fast limb movements. The audio elements in the "Tension" and "Fear" classes include high pitch, high tempo, high rhythmic variation, high volume, and a dissonant and complex harmony. The visual elements in the "Sad" class are closed arms, a face buried

in one's hands, hands touching the head, tears in eyes, a single person in a scene, a dark background, and slow-changing scenes. The "Relaxation" class includes ethnic music and is visually represented with natural scenes in slow motion and single-person performances with musical instruments. The acoustic components of the "Sad" and "Relaxation" classes include slow tempo, uniform harmonics, soft music, and low volume. The "Natural" class includes mixed characteristics from all five other classes. The data samples included in each class reflect diversity in music genres, culture and nation of origination, language, number of music sources, and mood. Five coworkers were involved in the construction of the new dataset.

The raw music video data needed to be processed in an acceptable form prior to being entered into the neural network. Our processing of the dataset of music, video, and facial expressions followed a number of steps for each individual data sample. The music network was trained on the real (magnitude) and imaginary (phase angle) components of the log magnitude spectrogram. The magnitude of the log Mel spectrogram was kept to one channel, while the phase angle representation was placed in another in order to preserve both.

For this work, a 30-second audio waveform x_i was converted into mono and then subsampled with a window size of 2048, a sampling rate of 22,050 Hz, and a time shift perimeter of 512 samples. The sampling rate varied for the slow–fast network, in which x_i in the slow path was sampled at a rate of 32,000 Hz, while the x_i in the fast path had a sampling rate of 8000 Hz. Fast-Fourier Transform (FFT) was then applied to each window to transform the x_i from a time-domain representation into a time–frequency (T-F) representation ($X_i(t, f)$). A total of 128 non-linear Mel scales that matched the human auditory system were selected from across the entire frequency spectrum. The log Mel spectrogram offers two advantages over waveform audio; first, it reduces the amount of data that the neural network needs to process compared with waveform representation, and second, it is correlated with human auditory perception and instrument frequency ranges [78].

The serially binned images were collected in a distributed way to preserve the temporal information of the entire video sequence, as shown in Figure 1. Each video was converted into several-frame sequences, $V_i = \{v\tau, v_2\tau, v_3\tau, \ldots, v_n\tau\}$, where τ represents equal time intervals in the video sequence. For each sample, τ changes according to the total number of frames n that were extracted. For the video/face network, 64 frames were taken in a distributed fashion. The video data were processed in a similar way to the audio data by varying the frame rate in the slow (eight frames) and fast (64 frames) branches of the slow–fast network.

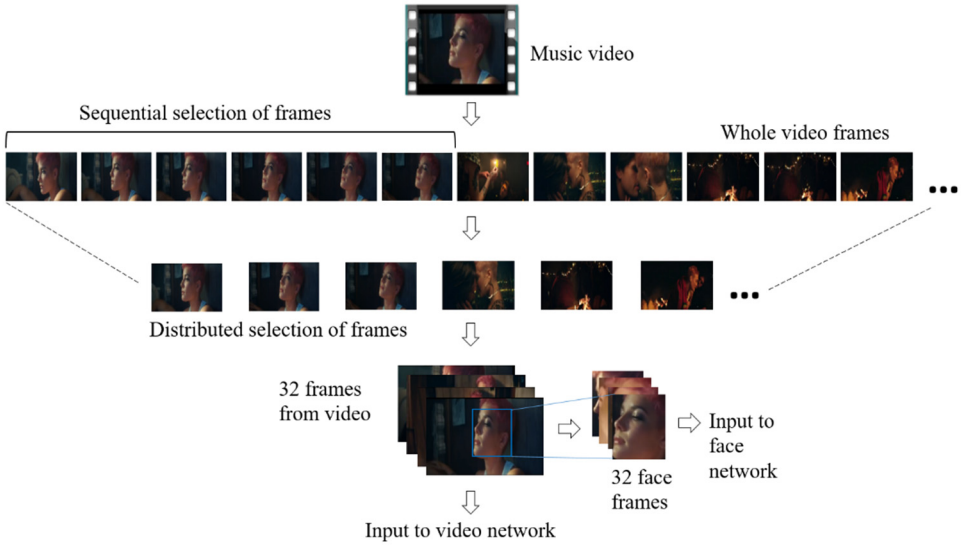


Figure 1. Input video processing by using the distributed selection of frames and faces.

For the face network input, face areas were detected in each video frame by using the cascade classifier feature of OpenCV (https://opencv.org/ accessed on 7 June 2021) 4.2.0 version. The images were cropped and resized for the network input. Music video frames may or may not have a face in them, or may depict more than one face. We chose music videos that contained at least one face. In each video, video frame v_t may contain one face $\{f_1\}$, more than one face $\{f_1, f_2, f_3, \ldots, f_m\}$, or no faces. After processing all of the video frames, we counted all of the frames containing faces. If the number of face frames was less than the face network input size, we repeated these frames until satisfying the requirement. Additional frames were discarded if the total face frames exceeded the neural network input. The repeated frames and discarded additional frames lost the temporal information of video sequences of faces. This is one reason for why our face network had a relatively small contribution to the overall decision making compared to the video and music networks.

After the preprocessing, the input for the audio network was $A_N = \sum_{i=0}^{N} X_i$; for the video network, it was $V_N = \sum_{i=0}^{N} V_i$, and for the face network, it was $F_N = \sum_{i=0}^{N} F_i$, where N is the data used in one batch. The multimodal input was the integrated form of each unimodal input.

4. Proposed Approach

4.1. Convolution Filter

Several networks were designed and integrated in a numbered way to find the optimal structure. The general 2D and 3D convolution filters were used in the music and video/face networks, respectively. Particularly in video processing, 3D convolution has been found to be better in capturing the spatial and motion information, but it exponentially increases the system's complexity. Popular 3D networks [79,80] have a great complexity and, hence, require large amounts of data for successful training.

In this paper, the complexity of 3D convolution was reduced by using separable filter and channel convolution. For the separable filter, the 3D convolution filter of size $n \times n \times n$ was divided into 2D space as $1 \times n \times n$ and $n \times n \times 1$. As illustrated in the right-most column of Figure 2, the 3D convolution filter was split into a 2D space filter with channel separation. The proposed convolution was an integrated form of separable channel convolution [81] (second column) and (2 + 1)D convolution [82] (third column). The idea of the separable filter and channel convolution was also used for the 2D audio network. The square filter of the 2D convolution was divided into a temporal filter ($1 \times n$) and a spatial filter ($n \times 1$), as in [83]. The channel size was reduced to one in the sequential block of the dense residual network for the separable channels. By using a novel separable channel and filter convolution, we drastically reduced the complexity and improved the system's performance for both the music and video networks.

4.2. Proposed Networks

We propose four basic architectures for the music, video, and face emotion networks. The architecture A_0 in Figure 3 only used 2D/3D convolution, and the other architectures (A_1 to A_3), which are shown in Figure 4, were designed with a separable 2D/3D channel and filter convolution. A_3 was a slow–fast network designed to capture the spatial and temporal information of audio/video. The basic architecture of our proposed network and a detailed view of each block are shown in Figures 3 and 4. In addition to the proposed networks, the well-known C3D network [79], which was trained on the sport-1M dataset [84], was also used for video and facial expression recognition. The numbers of filters in each of the five convolution layers (1–5) were 64, 128, 256, 512, and 512, respectively. The dimensions of the input network (height, width, channel, number of frames or depth) were equal to 112, 112, 3, and 32. The original C3D network was modified in a bottleneck layer with a dropout value of 0.2. The modified C3D network was the same as the one in [16], which helped us to make a fair comparison with this study based on the parameters and evaluation

scores. For a detailed look at the architecture, the reader is invited to refer to the original paper [79]. The pre-trained network for music video emotion classification was not found to be beneficial in terms of performance and complexity.

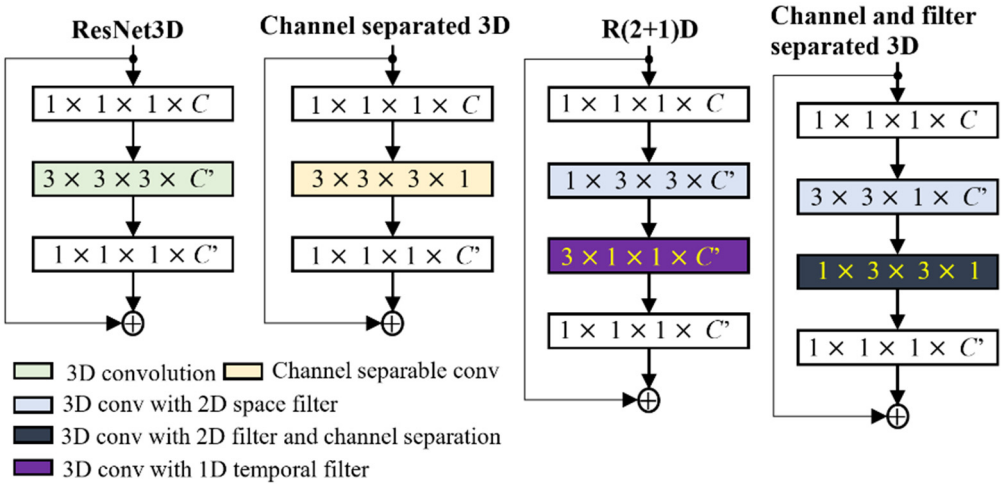


Figure 2. The 3D convolution and its variants in the residual block representation.

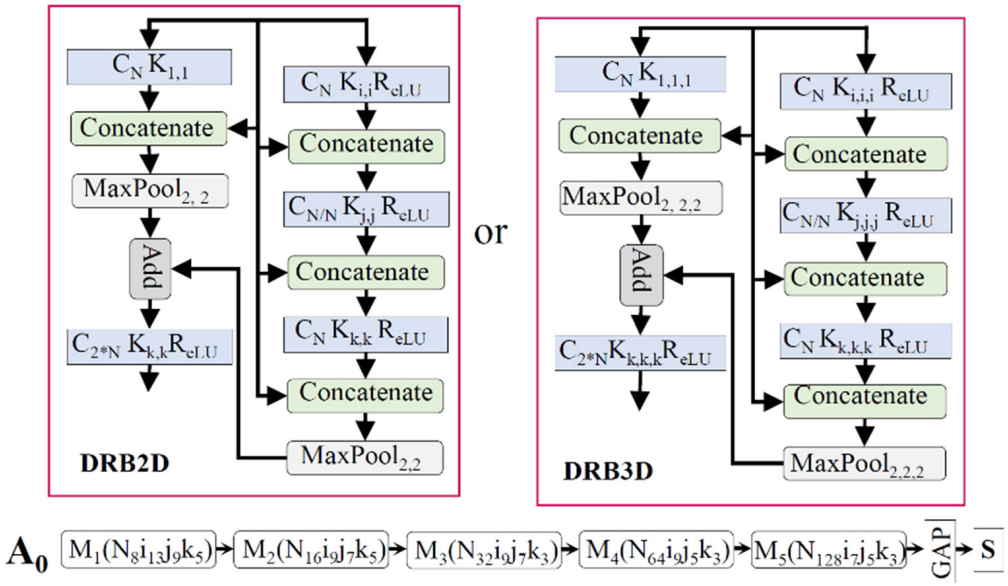


Figure 3. The block diagram of the proposed music video emotion classification network A_0 with the 2D/3D convolution. The acronyms DRB2DSC and DRB3DSC are detailed views of the dense residual block with the standard 2D and 3D convolution for the music network and the video/face network, respectively. The symbol A represents the network architecture, M represents the dense residual block defined in the detailed view, S represents the Softmax function, GAP is global average pooling, and MMTM is the multimodal transfer module. Similarly, the symbols N, i, j, and k with values in the lower case represent the values of the items, as illustrated in the respective detailed views.

The basic block of the unimodal architecture, which is shown in Figures 3 and 4, was further integrated for the multimodal representations. The feature information from multiple branches was merged to enhance top-level decision processing. A review [85] illustrated early, late, and mid-level fusion in multimodal affective computing. Differently from this research, our network learned information and integrated it after each block of the dense residual network by using MMTM. The MMTM and SE networks were used for information sharing and boosting during training. In the ablation study, we performed

an analysis of these blocks with several unimodal and multimodal architectures. At the decision level, all of the branch information was globally aggregated and computed for the class-wise probability. The Softmax function was used in the final layer of the neural network, which mapped the output nodes in a probability value range between 0 and 1. We used the categorical cross-entropy loss function for a one-hot vector target. This function was used to separately compute the binary cross-entropy for each class and then sum them up for the complete loss.

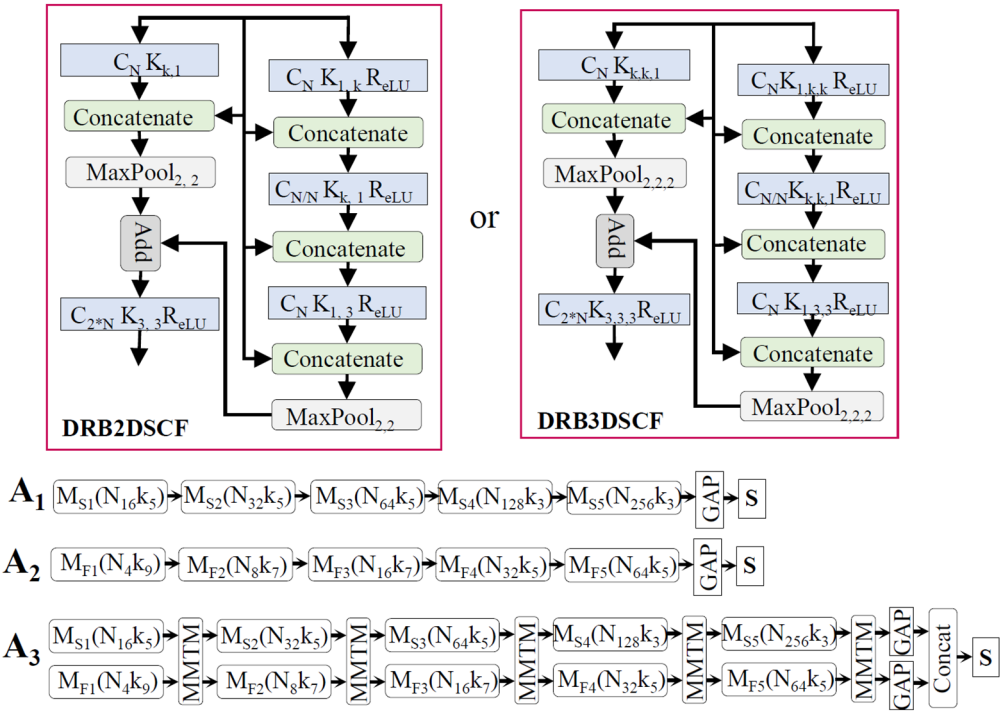


Figure 4. The block diagram of the proposed music video emotion classification network using slow (A_1), fast (A_2), or slow–fast network with MMTM (A_3) and with the separable channel and filter convolution. The acronyms DRB2DSCFC and DRB3DSCFC are the detailed views of the dense residual block with 2D and 3D separable channel and filter convolution for the music and video/face networks. The symbols have the same meanings as the symbols in Figure 3.

5. Experimental Results

To make this report understandable and comparable with those of different research groups, we consistently report our experimental data. Several networks are compared based on the evaluation score, complexity, and visual analysis by using a confusion matrix and a receiver operating characteristic (ROC) curve. We define accuracy as the probability of correct classification within a dataset. Accuracy indicates better evaluations in a balanced dataset. The F-measure is the harmonic mean of precision and recall; when the precision increases, the recall decreases, and vice versa. The F-score handles imbalanced data and provides a measure of the classifier's performance across different significance levels. A confusion matrix presents the number of correct and erroneous classifications by specifying erroneously categorized classes. A confusion matrix is a good option for reporting results in multi-class classification. The area under the ROC curve is a measure of how well a parameter can distinguish between a true positive and a true negative. An ROC is a probability curve that provides a measure of a classifier's performance in two-dimensional space. The area under the ROC curve (AUC) measures the degree of separability across multiple classification thresholds. These evaluation metrics were used to evaluate the effectiveness of our system for emotional classification in music videos. The data (https://zenodo.org/record/4542796#.YCxqhWgzaUk accessed on 7 June 2021) and code

(https://github.com/yagyapandeya/Supervised-Music-Video-Emotion-Classification accessed on 7 June 2021) used to produce these experiments are publicly available (in Supplementary Materials).

5.1. Results of the Unimodal Architectures

The unimodal architectures for music, video, and facial expressions were separately trained and tested. The testing dataset included 300 music video samples that were never used in the training process. These samples were equally distributed in six musical categories for a fair comparison. To measure the performance in terms of our evaluation metrics, the respective ground truth was provided for each test sample. An ablation study was performed to find an optimal network architecture that used both the unimodal and multimodal architectures. The system performance varied with the networks that used SE and MMTM. The SE network was proven to boost system performance. We only used the SE block in our unimodal architecture. In the A_0 unimodal architecture, the SE block proved to be beneficial. In the other networks, however, it was not found to be effective. The separation in the channel and convolution filter diversified the focal points of the network. The MMTM is an extended form of the SE block that allows more than one modality to share and enhance the information that it learns. The A_3 unimodal architectures were tested with the MMTM, and the results are illustrated in Tables 2–4. In the A_3 music architecture, the slow and fast paths did not properly synchronize due to their different sampling rates. Therefore, the music networks with or without MMTM showed poor performance. The A_3 face network showed poor performance because the temporal patterns of the face/video sequence were lost due to repeated or discarded frames. The MMTM block was found to be useful in the case of the A_3 video architecture. In this architecture, the temporal information was preserved and synchronized with spatial information during training. Each unimodal architecture was evaluated here in relation to multiple networks. The effects of the MMTM and SE blocks were evaluated for each network. These blocks placed increased complexity on the system, but were found to be more efficient in some cases.

Table 2. Test results using facial expression.

Model	Test Accuracy	F1-Score	ROC AUC Score	Parameters
C3D	0.3866	0.39	0.731	57,544,966
A_0 without SE (Face_A_0_noSE)	0.460	0.45	0.778	19,397,078
A_0 with SE (Face_A0_SE)	**0.516**	**0.51**	**0.820**	19,409,478
A_1 without SE (Face_A_1_noSE)	0.4933	0.46	0.810	11,876,982
A_1 with SE (Face_A_1_SE)	0.490	0.46	0.794	11,924,566
A_2 without SE (Face_A_2_noSE)	0.430	0.37	0.769	**845,670**
A_2 with SE (Face_A_2_SE)	0.403	0.37	0.755	849,406
A_3 without MMTM (Face_A_3_noMMTM)	0.449	0.42	0.781	24,083,846
A_3 with MMTM (Face_A_3_MMTM)	0.419	0.41	0.782	24,174,918

The bold number represents the highest evaluation score and lightweight network architecture (rightmost column).

Table 3. Test results using music information.

Model	Test Accuracy	F1-Score	ROC AUC Score	Parameters
A_0 without SE (Music_A_0_noSE)	0.5900	0.58	0.863	3,637,142
A_0 with SE (Music_A_0_SE)	0.5766	0.61	0.852	3,659,782
A_1 without SE (Music_A_1_noSE)	0.5366	0.51	0.859	3,946,949
A_1 with SE (Music_A_1_SE)	**0.6466**	**0.62**	**0.890**	3,994,533
A_2 without SE (Music_A_2_noSE)	0.6399	0.61	0.897	**261,297**
A_2 with SE (Music_A_2_SE)	0.6266	0.61	0.878	267,369
A_3 without MMTM (Music_A_3_noMMTM)	0.3166	0.22	0.635	4,208,240
A_3 with MMTM (Music_A_3_MMTM)	0.2433	0.17	0.610	7,941,004

The bold number represents the highest evaluation score and lightweight network architecture (rightmost column).

Table 4. Test results using video information.

Model	Test Accuracy	F1-Score	ROC AUC Score	Parameters
C3D	0.3266	0.19	0.723	57,544,966
A_0 without SE (Video_A_0_noSE)	0.4233	0.36	0.742	19,397,078
A_0 with SE (Video_A_0_SE)	0.4833	0.46	0.806	19,409,478
A_1 without SE (Video_A_1_noSE)	0.4099	0.39	0.754	11,876,982
A_1 with SE (Video_A_1_SE)	0.3666	0.35	0.736	11,922,518
A_2 without SE (Video_A_2_noSE)	0.3633	0.33	0.710	**845,670**
A_2 with SE (Video_A_2_SE)	0.3866	0.34	0.727	849,406
A_3 without MMTM (Vodeo_A_3_noMMTM)	0.4666	0.44	0.774	12,722,646
A_3 with MMTM (Vodeo_A_3_MMTM)	**0.5233**	**0.53**	**0.837**	24,174,918

The bold number represents the highest evaluation score and lightweight network architecture (rightmost column).

We evaluated the emotions in facial expressions based on the facial information of the music video actors. The face network performed poorly in relation to the music and video network because it could not deal with video frames with no faces or multiple faces. Uncertainty appeared in the system when it was presented with faces from an audience or supporting actors. Additionally, the system could not comprehend faces that were blurry and that were presented with low resolution, and this confusion reduced the performance. The emotional cues of the face, however, were still found to be important because they could boost the overall system's decisions. Table 2 shows the evaluation scores of the various face network architectures that we tested.

The musical analysis focused on the objective aspect of musicality. The neural network determined the changes in the spectral representations according to the emotional category. The unimodal architecture, which used only music information, performed the best compared to the face and video networks. The success of the network was related to the smooth changes in the musical patterns over time. Uncertainty in the music processing network, however, could arise due to the subjectivity of musical components and expressive techniques. We tested various music network architectures for emotion in music, and they are illustrated in Table 3. The A_3 music networks with and without MMTM had low performance rates because the spectral representation with varying sampling rates could not be synchronized. The single-branch network (2D, slow and fast) performed better with fewer parameters. Both positive and negative effects were found when using the squeeze-and-excitation blocks with the music networks.

The results of the video network were better than those of the face network, but not as good as those of the music network. The video network had a smooth temporal pattern that could not occur in the face network because a non-face frame would break the sequence. Compared to the music, the visual scenes abruptly changed according to time, which could affect the system's performance. Uncertainties in the video network could occur with user-generated videos, which may not have industry standards of recording, camera movement, and focus. We used the slow–fast network architecture (A_3) to capture the spatial and temporal information of videos with varying frame rates on each branch. The learned information of each branch was boosted and shared by using the MMTM block. Table 4 reports the various architectures and their scores on the evaluation metrics. Even though it had relatively large training parameters, the slow–fast network with MMTM performed the best when compared to the performance of the other architectures.

5.2. Result of the Multimodal Architecture

This study integrated several unimodal architectures for an efficient and optimal solution for emotion classification in music videos. The multimodal architecture was designed in two ways: using music and video information and using music, video, and facial information. We tested several combinations and obtained effective results.

Several combinations of music and video networks are possible; the best-performing multimodal architectures are shown in Table 5. While multimodal architectures that use the face network with audio or video were a possible network option, the contributions of the face network were minimal, and hence, the results are not discussed. The video and music architectures were found to be the dominant sources for overall prediction. With the multimodal video architecture, the MMTM block guided the two parallel branches of the slow–fast network by maintaining the proper synchronization of the learned information. The A_0 audio network outperformed the others in integration with video and achieved the same accuracy as our best multimodal architecture that used music, video, and facial information, as shown in Table 6.

Table 5. Test results using music and video information.

Model	Test Accuracy	F1-Score	ROC AUC Score	Parameters
Vodeo_A_3_MMTM + Music_A_0_noSE	**0.7400**	**0.71**	**0.938**	27,812,054
Vodeo_A_3_MMTM + Music_A_1_noSE	0.6733	0.66	0.919	28,121,861
Vodeo_A_3_MMTM + Music_A_2_noSE	0.6399	0.64	0.896	**24,436,209**

The bold number represents the highest evaluation score and lightweight network architecture (rightmost column).

Table 6. Integrated test results using music, video, and facial expressions.

Model	Test Accuracy	F1-Score	ROC AUC Score	Parameters
Vodeo_A_3_MMTM + Music_A_0_SE + Face_A_2_SE	**0.74000**	**0.73**	0.926	28,660,878
Vodeo_A_3_MMTM + Music_A_0_SE + Face_A_2_noSE	0.73333	0.72	**0.942**	28,589,478
Vodeo_A_3_MMTM + Music_A_0_noSE + Face_A_2_noSE	0.73333	0.71	0.939	28,657,718
Vodeo_A_3_MMTM + Music_A_0_SE + Face_A_1_noSE	0.6899	0.69	0.917	39,369,624
Vodeo_A_3_MMTM + Music_A_0_noSE + Face_A_1_noSE	0.71666	0.71	0.931	39,689,030
Video_A_1_noSE + Music_A_0_noSE + Face_A_2_noSE	0.69666	0.70	0.912	16,356,198
Video_A_2_noSE + Music_A_0_noSE + Face_A_2_noSE	0.68666	0.67	0.915	4,587,350
Video_A_2_noSE + Music_A_2_noSE + Face_A_2_noSE	0.610	0.59	0.873	**1,433,649**
Video_A_1_noSE + Music_A_1_noSE + Face_A_1_noSE	0.63666	0.63	0.860	19,432,869
Video_A_1_noSE + Music_A_1_noSE + Face_A_2_noSE	0.69999	0.69	0.925	11,942,805

The bold number represents the highest evaluation score and lightweight network architecture (rightmost column).

The integrated network of music, video, and facial expressions was trained in an end-to-end manner by using supervised data. Table 6 reports the integrated results for music, video, and facial expression information. The multimodal architecture that only used 3D convolution has extensive parameters, but the performance did not exceed that of the best unimodal music or video architecture. The slow–fast video network (A_3) with MMTM performed the best with the integrated architecture. The music network with 2D convolution (A_0) was found to be better than the network with the rectangular filter (A_1 and A_2). The first row of Table 6 shows our top-performing networks. The multimodal architecture using the A_2 network with music, video, and facial expressions (eighth row) had the lowest number of parameters, but the performance was even lower when using the best unimodal music architecture. The A_2 face network with SE was found to be effective in the integrated architecture with both music and video networks. For example, the A_2 face network with the A_1 music and video network (last row of Table 6) performed better than the A_1 face network (penultimate row). In the overall analysis, the visual clues that used face expression were found to be supportive for the classifier, with a small increment in network complexity.

5.3. Analysis Based on Visual Predictions

We validated the results of our experiments by using two visual evaluation methods: a confusion matrix and a ROC curve plot. The confusion matrix counted the number of samples in classes that were confused with each other. The confusion matrix in Figure 5 shows that the "Neutral" class was highly confusing for our classifier because it held data that were similar to those of more than one class. The classifier result showed confusion on the samples from the "Fear" and "Tension" classes because both classes held similar music structures (mostly rock and metal music). The rock and metal music samples also had common visual characteristics, such as angry facial expressions, dark backgrounds, and unique gestures and appearances. This number of commonalities confused the classifier. The "Sad" and "Relaxation" music videos had a similarly silent nature, so the classifier also confused these classes.

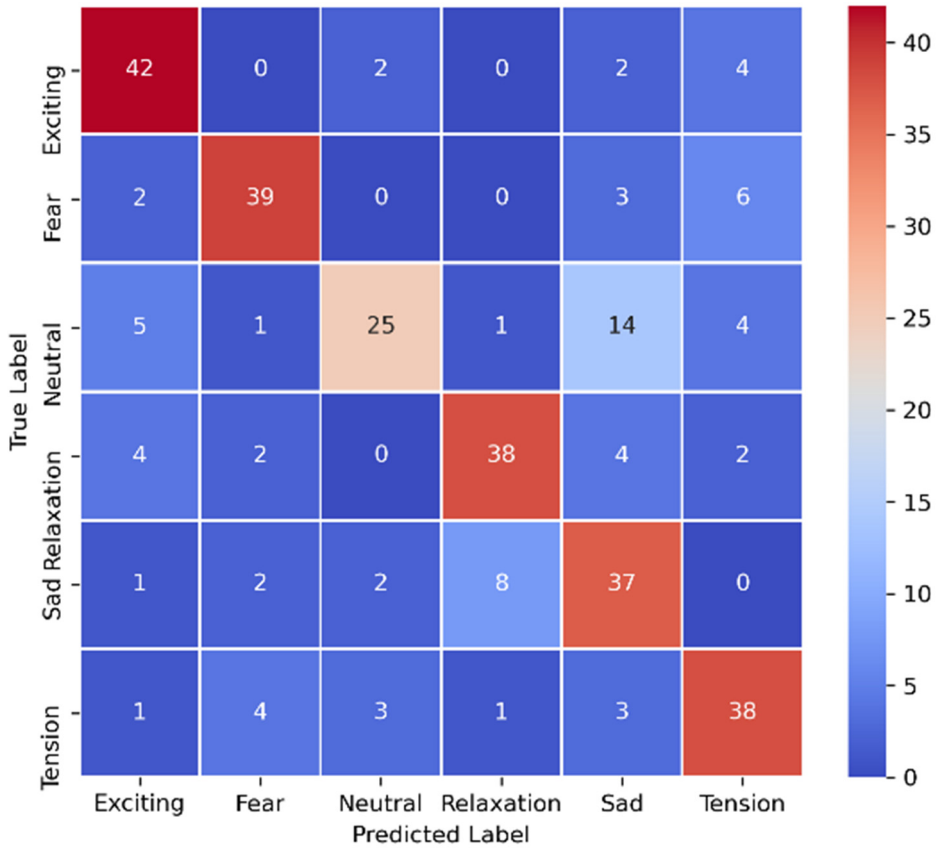


Figure 5. The confusion matrix using our best-performing multimodal architecture.

The ROC curves obtained when using our various multimodal architectures are shown in Figure 6. The multimodal architecture with music, video, and facial expression information performed the best. All of the classifiers showed similar ROC curves, with a small difference in the area under the curve (AUC) scores. Networks with similar performance and relatively fewer parameters obtained higher ROC-AUC scores.

Human emotions can be connected to each other, and these connections also appear in music videos. We analyzed the correlation of our six emotional categories in music videos. The class-wise probability was computed by using the sigmoid function at the end of the neural network. The class correlation results of test samples from each class are illustrated in Table 7. A single frame from each sample is provided for illustration. The results show that "Neutral" class carried various common features of the other remaining emotional classes. In addition, the samples from the "Fear" and "Tension" classes were found to be correlated with each other in this experiment, while the samples from the "Excited" and "Sad" classes were not found to be correlated.

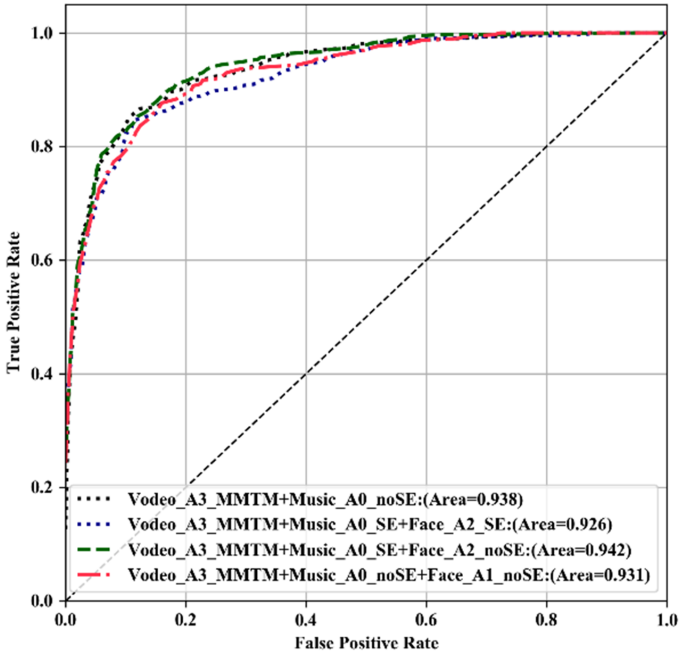
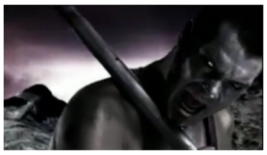


Figure 6. The ROC curve using several architectures for music, video and facial expressions.

Table 7. Test results using the best multimodal architecture.

Video Frame	Class-Wise Probability	Video Frame	Class-Wise Probability
 Ground truth: *Excited*	• Excited: 0.999981 • Relax: 0.884070 • Fear: 0.34791592 • Neutral: 0.251401 • Sad: 0.008204 • Tension: 0.006412	 Ground truth: *Neutral*	• Neutral: 0.999665 • Sad: 0.373862 • Relax: 0.301067 • Tension: 0.287840 • Excited: 0.109657 • Fear: 0.004261
 Ground truth: *Sad*	• Sad: 0.998427 • Relax: 0.741548 • Neutral: 0.402856 • Tension: 0.294600 • Excited: 0.027875 • Fear: 0.004099	 Ground truth: *Fear*	• Fear: 0.998849 • Tension: 0.945754 • Excited: 0.593985 • Neutral: 0.374574 • Sad: 0.003163 • Relax: 0.002293
 Ground truth: *Relaxation*	• Relax: 0.980475 • Excited: 0.973017 • Tension: 0.397293 • Neutral: 0.220959 • Sad: 0.177647 • Fear: 0.014223	 Ground truth: *Tension*	• Tension: 0.999754 • Neutral: 0.777341 • Fear: 0.4758184 • Sad: 0.03204632 • Relax: 0.03029701 • Excited: 0.008435

An emotional relationship between consumers and music videos was shown in [86–88] to be cross-culturally essential for promoting a song or album. We present the results of our best-performing network on two music video with billions of views on YouTube at the time of writing. The prediction results for *Despacito* (https://www.youtube.com/watch?v=kJQP7kiw5Fk) (accessed on 7 June 2021) (7.41B views) and *Gangnam Style* (https://www.youtube.com/watch?v=9bZkp7q19f0) (accessed on 7 June 2021) (4.10B views) are shown in Figure 7 on the left and right, respectively. The prediction results (in percentages) for each class are illustrated with individual curves depicted across 30-second increments. We used the sigmoid activation function to find the correlation of our six emotional categories. The

highest-activated class is illustrated with the highest score, and the second or third most probable classes are further down on the vertical axis. One video frame is illustrated at the bottom according to the time for illustration, but the musical components were also responsible for the final decision.

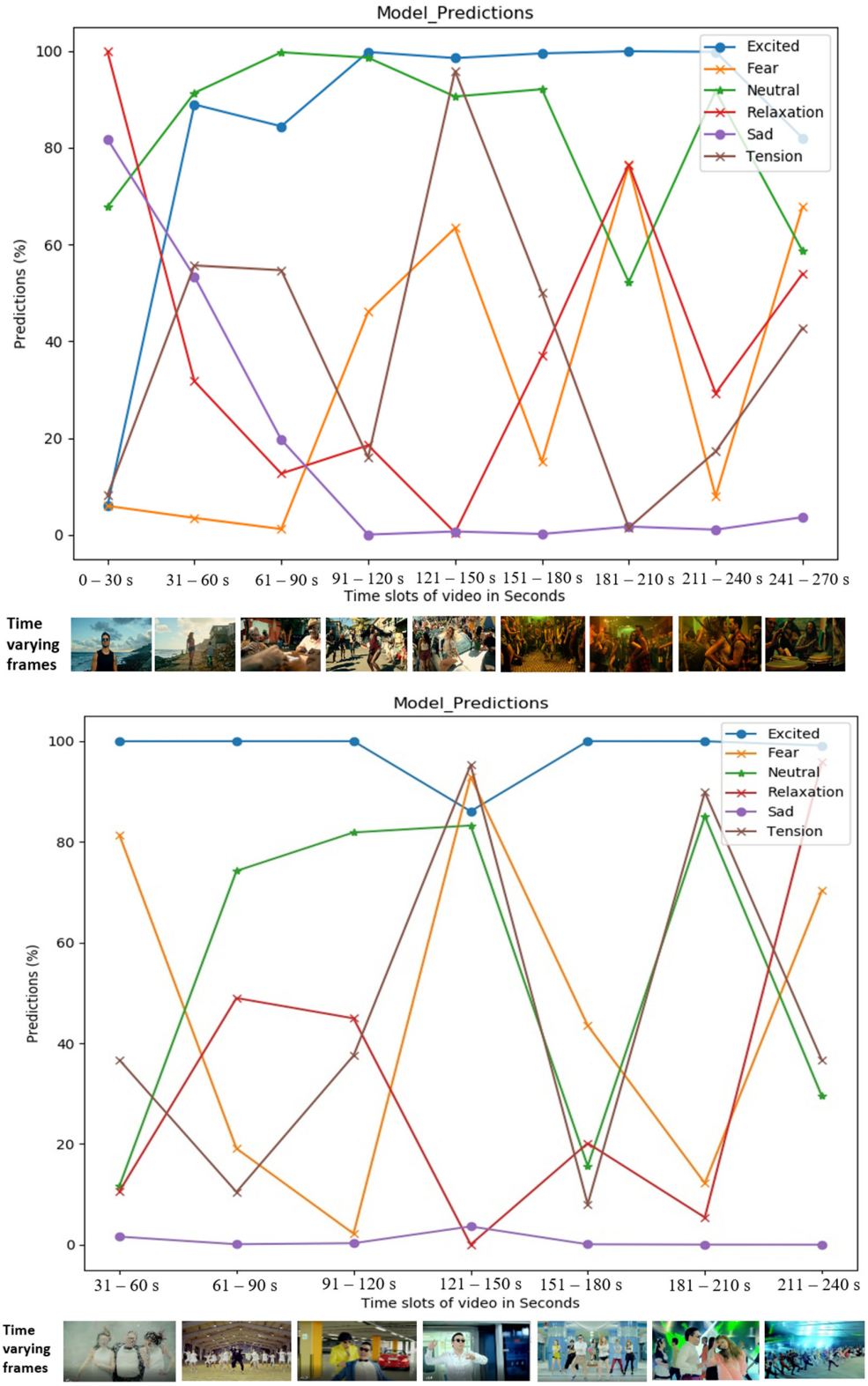


Figure 7. Time-synchronized emotional predictions for specific music videos. Prediction results for the *Luis_Fonsi* (**top**) and *Gangnam Style* (**bottom**) music videos.

5.4. Comparisons with Past Studies

Our research proposes a new deep learning method and prepares a new dataset for emotion classification in music videos. The past studies were conducted for affective computing on low- and mid-level audio and video feature classification using conventional classifiers, such as SVM. No study has implemented a deep learning method that can collectively use audio, video, and facial expression information in a single network with end-to-end training.

A previous study [16] attempted to use a similar music video dataset and a multimodal architecture (MVMM) to analyze music and video. The study implemented a two-stage process in which audio and video features were classified after transfer learning. The model resulted in a good evaluation score, as their test samples were taken from similar training samples. However, the pre-trained C3D network used in [16] had a relatively large number of parameters to which even more parameters were added by our combination of this network with other audio networks (a more detailed comparison of the performance and complexity of our network with those of the C3D network appears in Section 5). Our model reduced the network complexity and was evaluated with a more diverse set of test samples.

Some studies have proposed the use of a diverse dataset for affective computing of music videos. One study concerning emotional analysis of music that used a recurrent neural network with an SVM on top [28] achieved a classification accuracy of 0.542 with the LastFM dataset. CNN-based music emotion classification [32] achieved the highest F1-scores when using the CAL500 (0.534) and CAL500 exp (0.709) datasets with 18 emotion tags. The authors of [25] used an SVM for low-level feature classification of music, and ultimately attained the highest F1-score of 0.764. The results of this study were based on this research team's self-developed dataset, which consisted of 900 audio clips and associated subjective annotations that were applied consistently with Russell's emotion quadrants.

In Table 8, we quantitatively compare the current study with past related research. The proposed classifier could not outperform the quantitative results, but they were qualitatively robust because the networks were trained on a relatively large data sample with three sources of input. Hence, the network's capabilities were more diverse and applicable for real-world applications. The results of the visual analysis support this claim.

Table 8. Comparison with past studies.

Method	Dataset	Data Type	Emotion Class	Score
RNN [25]	LastFM	Music	4	0.542 (Accuracy)
CNN [22]	CAL500	Music	18	0.534 (F1-score)
	CAL500 exp	Music	18	0.709 (F1-Score)
SVM [21]	Own	Music	4	0.764(F1-Score)
GMM [56]	DEAP120	Music and video	8	0.90 (Accuracy)
CLR [50]	CAL500	Music and video	18	0.744 (Accuracy)
MM [50]	Own	Music and video	6	0.88 (F1-Score)
Our	Own	Music and video	6	0.71 (F1-Score)
Our	Own	Music, video and Face	6	0.73 (F1-Score)

A number of possible multimodal architectures that use music-related audio and video information have also been studied for affective computing. The study in [66] used low-level video features (lighting key, shot boundary, color, motion) and audio features (zero crossing, MFCC, delta MFCC) for emotion classification by using Gaussian mixture model (GMM) classifiers. This research team used the DEAP120 dataset with an eight-class category and reached 90% accuracy. The team behind [60] used 140 annotated music video

samples from the CAL500 dataset in a model that had the highest level of accuracy when used on audio and video with additional optical flow features (74.4%) or with audio and ImageNet features, as well as the Calibrated Label Ranking (CLR) classifier (72.24%). Two additional studies [60,66] were performed with a high accuracy despite the limited number of data samples that they were tested against. In each of these cases, emotion annotation was provided by the video or audio audience, a fact that differentiates our labeled video clips. Moreover, these methods used conventional emotion classification methods, unlike our end-to-end deep neural network, which accounts for multimodal information.

In this research, first, we analyzed the unimodal representations of music, video, and facial expression information from music videos by using end-to-end training. The unimodal architectures were further integrated into a multimodal architecture to develop a robust and optimal classifier. Compared to past research, our dataset is more diversified; the networks were trained and tested on real data domains and not on features. We reduced the system's complexity and enhanced the performance of the architecture by using novel convolution and information-boosting methods. The results were statistically evaluated by using various evaluation metrics.

6. Conclusions

Affective computing enables AI systems to interpret human emotions. This area of computing is inherently interdisciplinary, though the analysis of emotions in music videos remains a particularly unexplored area within the field of computer engineering. Our system classifies music videos by using a dataset that was introduced for supervised training. Several unimodal and multimodal architectures have been proposed to analyze music, video, and facial expressions from scratch. Our proposed architectures, including the slow–fast network, were designed to use 2D and 3D convolution, as well as a novel separable channel and filter convolution. Our best multimodal architecture achieved 74.00% accuracy, an F1-score of 0.73, and an area under the curve (AUC) score of 0.923.

Future researchers have space for improvement in terms of the performance, the dataset, the emotion representation framework, and the evaluation measures. In this study, we only included music, video, and facial expression networks. Music lyrics are another vital source of information that can be integrated in further studies for more accurate affective computing.

Supplementary Materials: The music video dataset used in this paper is available online at https://zenodo.org/record/4542796#.YCxqhWgzaUk (accessed on 7 June 2021) and code for music video emotion using multimodal system of music, video and face expression is available online at https://github.com/yagyapandeya/Supervised-Music-Video-Emotion-Classification (accessed on 7 June 2021).

Author Contributions: The first author, Y.R.P., contributed to the entire project, which included the conceptualization, methodology, software development, validation, analysis, data curation, and writing of the original drift. The corresponding author, J.L., contributed with the funding acquisition, conceptualization, and supervision. The second author, B.B., mainly helped in the review and editing, data curation, and resource management. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) under the Development of AI for Analysis and Synthesis of Korean Pansori NRF-2021R1A2C2006895 Project.

Acknowledgments: We would also like to express our gratitude to the editors of the Writing Center at Jeonbuk National University for their skilled English-language assistance.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MER	Music Emotion Recognition
DNN	Deep Neural Network
CNN	Convolutional Neural Network
GMM	Gaussian mixture model
SVM	Support Vector Machine
CLR	Calibrated Label Ranking
MVMM	Music Video Multi-Modal
C3D	Convolutional 3 Dimensional
CAL500	Computer Audition Lab 500-song
DEAP120	Database for Emotion Analysis using Physiological signals with 120 samples
MMTM	Multimodal Transfer Module
SE	Squeeze-and-Excitation
MFCC	Mel Frequency Cepstral Coefficient
ROC	Receiver Operation Characteristics
AUC	Area Under Curve
GAP	Global Average Pooling
FFT	Fast Fourier Transform
T-F	Time-Frequency
2/3D	2/3 Dimension
DRB2DSC	Dense Residual Block 2D Standard Convolution
DRB3DSC	Dense Residual Block 3D Standard Convolution
DRB2DSCFC	Dense Residual Block 2D with Separable Channel and Filter Convolution
DRB3DSCFC	Dense Residual Block 3D with Separable Channel and Filter Convolution

References

1. Yang, Y.H.; Chen, H.H. Machine Recognition of Music Emotion: A Review. *ACM Trans. Intell. Syst. Technol.* **2012**. [CrossRef]
2. Juslin, P.N.; Laukka, P. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *J. New Music Res.* **2004**, *33*, 217–238. [CrossRef]
3. Elvers, P.; Fischinger, T.; Steffens, J. Music Listening as Self-enhancement: Effects of Empowering Music on Momentary Explicit and Implicit Self-esteem. *Psychol. Music* **2018**, *46*, 307–325. [CrossRef]
4. Raglio, A.; Attardo, L.; Gontero, G.; Rollino, S.; Groppo, E.; Granieri, E. Effects of Music and Music Therapy on Mood in Neurological Patients. *World J. Psychiatry* **2015**, *5*, 68–78. [CrossRef] [PubMed]
5. Patricia, E.B. Music as a Mood Modulator. Retrospective Theses and Dissertations, 1992, 17311. Available online: https://lib.dr.iastate.edu/rtd/17311 (accessed on 7 June 2017).
6. Eerola, T.; Peltola, H.R. Memorable Experiences with Sad Music—Reasons, Reactions and Mechanisms of Three Types of Experiences. *PLoS ONE* **2016**, *11*, e0157444. [CrossRef] [PubMed]
7. Bogt, T.; Canale, N.; Lenzi, M.; Vieno, A.; Eijnden, R. Sad Music Depresses Sad Adolescents: A Listener's Profile. *Psychol. Music* **2019**, *49*, 257–272. [CrossRef]
8. Pannese, A.; Rappaz, M.A.; Grandjean, G. Metaphor and Music Emotion: Ancient Views and Future Directions. *Conscious. Cogn.* **2016**, *44*, 61–71. [CrossRef] [PubMed]
9. Siles, I.; Segura-Castillo, A.; Sancho, M.; Solís-Quesada, R. Genres as Social Affect: Cultivating Moods and Emotions through Playlists on Spotify. *Soc. Media Soc.* **2019**, *5*, 2056305119847514. [CrossRef]
10. Schriewer, K.; Bulaj, G. Music Streaming Services as Adjunct Therapies for Depression, Anxiety, and Bipolar Symptoms: Convergence of Digital Technologies, Mobile Apps, Emotions, and Global Mental Health. *Front. Public Health* **2016**, *4*, 217. [CrossRef]
11. Pandeya, Y.R.; Kim, D.; Lee, J. Domestic Cat Sound Classification Using Learned Features from Deep Neural Nets. *Appl. Sci.* **2018**, *8*, 1949. [CrossRef]
12. Pandeya, Y.R.; Bhattarai, B.; Lee, J. Visual Object Detector for Cow Sound Event Detection. *IEEE Access* **2020**, *8*, 162625–162633. [CrossRef]
13. Pandeya, Y.R.; Lee, J. Domestic Cat Sound Classification Using Transfer Learning. *Int. J. Fuzzy Log. Intell. Syst.* **2018**, *18*, 154–160. [CrossRef]
14. Pandeya, Y.R.; Bhattarai, B.; Lee, J. Sound Event Detection in Cowshed using Synthetic Data and Convolutional Neural Network. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 21–23 October 2020; pp. 273–276.
15. Bhattarai, B.; Pandeya, Y.R.; Lee, J. Parallel Stacked Hourglass Network for Music Source Separatio. *IEEE Access* **2020**, *8*, 206016–206027. [CrossRef]
16. Pandeya, Y.R.; Lee, J. Deep Learning-based Late Fusion of Multimodal Information for Emotion Classification of Music Video. *Multimed. Tools Appl.* **2020**, *80*, 2887–2905. [CrossRef]

17. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October 2019–2 November 2019.
18. Joze, H.R.V.; Shaban, A.; Iuzzolino, M.L.; Koishida, K. MMTM: Multimodal Transfer Module for CNN Fusion. In Proceedings of the CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
19. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
20. Lopes, P.; Liapis, A.; Yannakakis, G.N. Modelling Affect for Horror Soundscapes. *IEEE Trans. Affect. Comput.* **2019**, *10*, 209–222. [CrossRef]
21. Naoki, N.; Katsutoshi, I.; Hiromasa, F.; Goto, M.; Ogata, T.; Okuno, H.G. A Musical Mood Trajectory Estimation Method Using Lyrics and Acoustic Features. In Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies, Scottsdale, AZ, USA, 28 November 2011–1 December 2011; pp. 51–56.
22. Song, Y.; Dixon, S.; Pearce, M. Evaluation of Musical Features for Music Emotion Classification. In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, 8–12 October 2012; pp. 523–528.
23. Lin, C.; Liu, M.; Hsiung, W.; Jhang, J. Music Emotion Recognition Based on Two-level Support Vector Classification. In Proceedings of the 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju Island, Korea, 10–13 July 2016; pp. 375–389.
24. Han, K.M.; Zin, T.; Tun, H.M. Extraction of Audio Features for Emotion Recognition System Based on Music. *Int. J. Sci. Technol. Res.* **2016**, *5*, 53–56.
25. Panda, R.; Malheiro, R.; Paiva, R.P. Novel Audio Features for Music Emotion Recognition. *IEEE Trans. Affect. Comput.* **2020**, *11*, 614–626. [CrossRef]
26. Aljanaki, A.; Yang, Y.H.; Soleymani, M. Developing a Benchmark for Emotional Analysis of Music. *PLoS ONE* **2017**, *12*, e0173392. [CrossRef]
27. Malik, M.; Adavanne, A.; Drossos, K.; Virtanen, T.; Ticha, D.; Jarina, R. Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. *arXiv* **2017**, arXiv:1706.02292v1. Available online: https://arxiv.org/abs/1706.02292 (accessed on 7 June 2017).
28. Jakubik, J.; Kwaśnicka, H. Music Emotion Analysis using Semantic Embedding Recurrent Neural Networks. In Proceedings of the 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Gdynia, Poland, 3–5 July 2017; pp. 271–276.
29. Liu, X.; Chen, Q.; Wu, X.; Yan, L.; Yang, L. CNN Based Music Emotion Classification. *arXiv* **2017**, arXiv:1704.05665. Available online: https://arxiv.org/abs/1704.05665 (accessed on 19 April 2017).
30. Tsunoo, E.; Akase, T.; Ono, N.; Sagayama, S. Music mood classification by rhythm and bass-line unit pattern analysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 265–268.
31. Turnbull, D.; Barrington, L.; Torres, D.; Lanckriet, G. Towards musical query-by-semantic description using the cal500 data set. In Proceedings of the ACM SIGIR, Amsterdam, The Netherlands, 23–27 July 2007; pp. 439–446.
32. Li, S.; Huang, L. Music Emotions Recognition Based on Feature Analysis. In Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 13–15 October 2018; pp. 1–5.
33. Wang, S.; Wang, J.; Yang, Y.; Wang, H. Towards time-varying music auto-tagging based on cal500 expansion. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14–18 July 2014; pp. 1–6.
34. Berardinis, J.; Cangelosi, A.; Coutinho, E. The Multiple Voices of Music Emotions: Source Separation for Improving Music Emotion Recognition Models and Their Interpretability. In Proceedings of the ISMIR 2020, Montréal, QC, Canada, 11–16 October 2020.
35. Chaki, S.; Doshi, P.; Bhattacharya, S.; Patnaik, P. Explaining Perceived Emotions in Music: An Attentive Approach. In Proceedings of the ISMIR 2020, Montréal, QC, Canada, 11–16 October 2020.
36. Orjesek, R.; Jarina, R.; Chmulik, M.; Kuba, M. DNN Based Music Emotion Recognition from Raw Audio Signal. In Proceedings of the 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019; pp. 1–4.
37. Choi, W.; Kim, M.; Chung, J.; Lee, D.; Jung, S. Investigating U-nets with Various Intermediate blocks for Spectrogram-Based Singing Voice Separation. In Proceedings of the ISMIR2020, Montréal, QC, Canada, 11–16 October 2020.
38. Yin, D.; Luo, C.; Xiong, Z.; Zeng, W. Phasen: A phase-and-harmonics-aware speech enhancement network. *arXiv* **2019**, arXiv:1911.04697. Available online: https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1773.html (accessed on 12 November 2019).
39. Takahashi, N.; Agrawal, P.; Goswami, N.; Mitsufuji, Y. Phasenet: Discretized phase modeling with deep neural networks for audio source separation. *Interspeech* **2018**, 2713–2717. [CrossRef]
40. Zhang, H.; Xu, M. Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 629–633.

41. Xu, B.; Fu, Y.; Jiang, Y.; Li, B.; Sigal, L. Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization. *IEEE Trans. Affect. Comput.* **2018**, *9*, 255–270. [CrossRef]
42. Tu, G.; Fu, Y.; Li, B.; Gao, J.; Jiang, Y.; Xue, X. A Multi-Task Neural Approach for Emotion Attribution, Classification, and Summarization. *IEEE Trans. Multimed.* **2020**, *22*, 148–159. [CrossRef]
43. Lee, J.; Kim, S.; Kiim, S.; Sohn, K. Spatiotemporal Attention Based Deep Neural Networks for Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1513–1517.
44. Sun, M.; Hsu, S.; Yang, M.; Chien, J. Context-aware Cascade Attention-based RNN for Video Emotion Recognition. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–6.
45. Xu, B.; Zheng, Y.; Ye, H.; Wu, C.; Wang, H.; Sun, G. Video Emotion Recognition with Concept Selection. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 406–411.
46. Irie, G.; Satou, T.; Kojima, A.; Yamasaki, T.; Aizawa, K. Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification. *IEEE Trans. Multimedia* **2010**, *12*, 523–535. [CrossRef]
47. Mo, S.; Niu, J.; Su, Y.; Das, S.K. A Novel Feature Set for Video Emotion Recognition. *Neurocomputing* **2018**, *291*, 11–20. [CrossRef]
48. Kaya, H.; Gürpınar, F.; Salah, A.A. Video-based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion. *Image Vis. Comput.* **2017**, *65*, 66–75. [CrossRef]
49. Li, H.; Kumar, N.; Chen, R.; Georgiou, P. A Deep Reinforcement Learning Framework for Identifying Funny Scenes in Movies. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3116–3120.
50. Ekman, P.; Friesen, W.V. Constants Across Cultures in the Face and Emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124.
51. Pantic, M.; Rothkrantz, L.J.M. Automatic Analysis of Facial Expressions: The State of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]
52. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
53. Majumder, A.; Behera, L.; Subramanian, V.K. Automatic Facial Expression Recognition System Using Deep Network-Based Data Fusion. *IEEE Trans. Cybern.* **2018**, *48*, 103–114. [CrossRef]
54. Kuo, C.; Lai, S.; Sarkis, M. A Compact Deep Learning Model for Robust Facial Expression Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2202–22028.
55. Nanda, A.; Im, W.; Choi, K.S.; Yang, H.S. Combined Center Dispersion Loss Function for Deep Facial Expression Recognition. *Pattern Recognit. Lett.* **2021**, *141*, 8–15. [CrossRef]
56. Tao, F.; Busso, C. End-to-End Audiovisual Speech Recognition System with Multitask Learning. *IEEE Trans. Multimed.* **2021**, *23*, 1–11. [CrossRef]
57. Eskimez, S.E.; Maddox, R.K.; Xu, C.; Duan, Z. Noise-Resilient Training Method for Face Landmark Generation from Speech. In Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing, Los Altos, CA, USA, 16 October 2019; Volume 28, pp. 27–38. [CrossRef]
58. Zeng, H.; Wang, X.; Wu, A.; Wang, Y.; Li, Q.; Endert, A.; Qu, H. EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 927–937. [CrossRef] [PubMed]
59. Seanglidet, Y.; Lee, B.S.; Yeo, C.K. Mood prediction from facial video with music "therapy" on a smartphone. In Proceedings of the 2016 Wireless Telecommunications Symposium (WTS), London, UK, 18–20 April 2016; pp. 1–5.
60. Kostiuk, B.; Costa, Y.M.G.; Britto, A.S.; Hu, X.; Silla, C.N. Multi-label Emotion Classification in Music Videos Using Ensembles of Audio and Video Features. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 517–523.
61. Acar, E.; Hopfgartner, F.; Albayrak, S. Understanding Affective Content of Music Videos through Learned Representations. In Proceedings of the International Conference on Multimedia Modeling, Dublin, Ireland, 10–18 January 2014.
62. Ekman, P. *Basic Emotions in Handbook of Cognition and Emotion*; Wiley: Hoboken, NJ, USA, 1999; pp. 45–60.
63. Russell, J.A. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
64. Thayer, R.E. *The Biopsychology of Mood and Arousal*; Oxford University Press: Oxford, UK, 1989.
65. Plutchik, R. *A General Psychoevolutionary Theory of Emotion in Theories of Emotion*, 4th ed.; Academic Press: Cambridge, MA, USA, 1980; pp. 3–33.
66. Skodras, E.; Fakotakis, N.; Ebrahimi, T. Multimedia Content Analysis for Emotional Characterization of Music Video Clips. *EURASIP J. Image Video Process.* **2013**, *2013*, 26.
67. Gómez-Cañón, J.S.; Cano, E.; Herrera, P.; Gómez, E. Joyful for You and Tender for Us: The Influence of Individual Characteristics and Language on Emotion Labeling and Classification. In Proceedings of the ISMIR 2020, Montréal, QC, Canada, 11–16 October 2020.
68. Eerola, T.; Vuoskoski, J.K. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music* **2011**, *39*, 18–49. [CrossRef]
69. Makris, D.; Kermanidis, K.L.; Karydis, I. The Greek Audio Dataset. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Rhodes, Greece, 19–21 September 2014.

70. Aljanaki, A.; Wiering, F.; Veltkamp, R.C. Studying emotion induced by music through a crowdsourcing game. *Inf. Process. Manag.* **2016**, *52*, 115–128. [CrossRef]
71. Yang, Y.H.; Lin, Y.C.; Su, Y.F.; Chen, H.H. A Regression Approach to Music Emotion Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 448–457. [CrossRef]
72. Livingstone, S.R.; Russo, R.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
73. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-Aware Emotion Recognition Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
74. Malandrakis, N.; Potamianos, A.; Evangelopoulos, G.; Zlatintsi, A. A supervised approach to movie emotion tracking. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 2376–2379.
75. Baveye, Y.; Dellandrea, E.; Chamaret, C.; Chen, L. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Trans. Affect. Comput.* **2015**, *6*, 43–55. [CrossRef]
76. Yang, Y.H.; Chen, H.H. *Music Emotion Recognition*; CRC Press: Boca Raton, FL, USA, 2011.
77. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *arXiv* **2021**, arXiv:2004.07780v4. Available online: https://arxiv.org/abs/2004.07780 (accessed on 16 April 2020). [CrossRef]
78. CJ-Moore, B. *An Introduction to the Psychology of Hearing*; Brill: Leiden, The Netherlands, 2012.
79. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
80. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. *arXiv* **2018**, arXiv:1705.07750v3.
81. Du, T.; Heng, W.; Lorenzo, T.; Matt, F. Video Classification with Channel-Separated Convolutional Networks. *arXiv* **2019**, arXiv:1904.02811v4. Available online: https://arxiv.org/abs/1904.02811 (accessed on 4 April 2019).
82. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
83. Pons, J.; Lidy, T.; Serra, X. Experimenting with musically motivated convolutional neural networks. In Proceedings of the 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), Bucharest, Romania, 15–17 June 2016; pp. 1–6.
84. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
85. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
86. Morris, J.D.; Boone, M.A. The Effects of Music on Emotional Response, Brand Attitude, and Purchase Intent in an Emotional Advertising Condition. *Adv. Consum. Res.* **1998**, *25*, 518–526.
87. Park, J.; Park, J.; Park, J. The Effects of User Engagements for User and Company Generated Videos on Music Sales: Empirical Evidence from YouTube. *Front. Psychol.* **2018**, *9*, 1880. [CrossRef]
88. Abolhasani, M.; Oakes, S.; Oakes, H. Music in advertising and consumer identity: The search for Heideggerian authenticity. *Mark. Theory* **2017**, *17*, 473–490. [CrossRef]