EmoMV: Affective Music-Video Correspondence Learning Datasets for Classification and Retrieval

Ha Thi Phuong Thao[a], Gemma Roig[b], Dorien Herremans[a,*]

[a]*Singapore University of Technology and Design, 8 Somapah Rd, Singapore 48737*
[b]*Goethe University Frankfurt, Department of Computer Science, Robert-Mayer-Str. 11-15, 60323 Frankfurt, Germany*

Abstract

Studies in affective audio-visual correspondence learning require ground-truth data to train, validate, and test models. The number of available datasets together with benchmarks, however, is still limited. In this paper, we create a collection of three datasets (called EmoMV) for affective correspondence learning between music and video modalities. The first two datasets (called EmoMV-A, and EmoMV-B, respectively) are constructed by making use of music video segments from other available datasets. The third one called EmoMV-C is created from music videos that we self-collected from YouTube. The music-video pairs in our datasets are annotated as matched or mismatched in terms of the emotions they are conveying. The emotions are annotated by humans in the EmoMV-A dataset, while in the EmoMV-B and EmoMV-C datasets they are predicted using a pretrained deep neural network. A user study is carried out to evaluate the accuracy of the "matched" and "mismatched" labels offered in the EmoMV dataset collection. In addition to creating three new datasets, a benchmark deep neural network model for binary affective music-video correspondence classification is also proposed. This proposed benchmark model is then modified to adapt to affective music-video retrieval. Extensive experiments are carried out on all three datasets of the EmoMV collection. Experimental results demonstrate that our proposed model outperforms state-of-the-art ap-

[*] Corresponding author
Email address: `dorien_herremans@sutd.edu.sg` (Dorien Herremans)

Preprint submitted to Information Fusion *August 6, 2022*

proaches on both the binary classification and retrieval tasks. We envision that our newly created dataset collection together with the proposed benchmark models will facilitate advances in affective computing research.

Keywords: Multi-task learning deep neural networks, affective audio-visual correspondence learing, emotion-based matching, affective music-video retrieval, EmoMV dataset collection, affective computing.

1. Introduction

Audio-visual correspondence learning aims to discover the global semantic link between visual and audio modalities [1]. The first audio-visual correspondence learning task was introduced in [2] as a binary classification task to classify whether a pair of an audio clip and a video frame matches (also called "positive" if they are extracted at the same time from the same video) or mismatches (i.e. "negative" if they come from different videos). Since then, audio-visual correspondence learning has been further investigated in many other studies [2, 3, 4, 5, 6, 7, 8], which often use various definitions of matched and mismatched pairs of audio and visual modalities. Some notable studies are audio-visual localization [9, 10, 11] (which deals with localizing the objects in videos that make sound), theme correspondence between videos and audio [12], faces and voices correspondence [13].

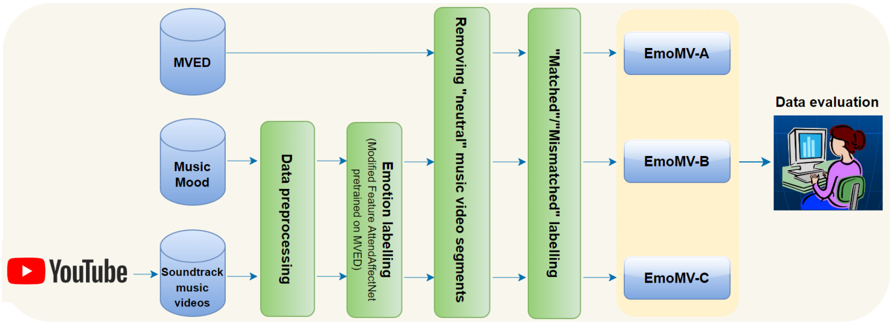

MVED
Music Mood
YouTube
Soundtrack music videos
Data preprocessing
Emotion labelling
(Modified Feature AttendAffectNet pretrained on MVED)
Removing "neutral" music video segments
"Matched"/"Mismatched" labelling
EmoMV-A
EmoMV-B
EmoMV-C
Data evaluation

Figure 1: Visualization of the process of creating the EmoMV dataset collection.

Both music and video can deliver rich meanings, and they are commonly used to evoke emotions [14]. By adding music to videos that convey similar emotions, we may perceive emotions in videos more vividly [4]. Such a combination could be useful for various applications, such as recommendation systems that recommend music to videos such that they evoke similar emotions. This may, for instance, allow filmmakers to add music to their videos to convey specific emotions. It may also allow advertisers to search for the perfect tune to accompany their videos to elicit a desired emotion in viewers. Although there are some studies on the correspondence between music and videos in terms of emotions [5, 4], this task remains a challenge.

The study on *affective* audio-visual correspondence learning, i.e. learning the correspondence or matching in terms of emotions between audio and visual modalities, has surprisingly not received a lot of attention. Most datasets used in studies on the affective correspondence between music and visual modalities are not publicly available due to copyright restrictions. This makes it difficult to improve on existing models, compare the performance using benchmarks, and in general advance the field. Therefore, we believe that the construction of open datasets for this task is a necessity. Compared to affective correspondence learning between music and videos, the music video emotion classification task has been studied more [15, 16, 17], and there are more available datasets. To tackle the problem of limited available data for affective audio-visual correspondence learning, we create a collection of three datasets called Emotion-based Music Video Matching (EmoMV) from various sources (including music videos collected from YouTube as well as other available datasets that were originally used for emotion classification). In the EmoMV collection, a pair of music and video is considered as having true correspondence (i.e. matched) if they carry similar emotional information, otherwise, we label them as having false correspondence (i.e. mismatched). This definition is inspired by the one mentioned in [18] for affective correspondence learning between music and images. The first dataset (called EmoMV-A) includes $4,914$ music video segments ($4,110$ for training, 556 for validation, and 248 for testing) with a duration of 30 seconds each. This dataset is created by making use of music video segments (annotated with an emotion label from categories including exciting, fearful, tense, sad, and relaxing) from the Music Video Emotion Dataset (MVED) [19]. The second dataset (called EmoMV-B) is created from music video segments selected from the Music Mood dataset of the AudioSet ontology [20], which was originally used for the audio classification task. The EmoMV-B dataset includes 616 music video segments (496 for training, and 120 for validation) with a duration of 10 seconds each. The third dataset (called EmoMV-C) includes 456 music video segments (360 for training and 96 for validation) with a duration of 30 seconds each. These music video segments are split from music videos of songs featured in movies (i.e. soundtrack music videos) that we self-collected from YouTube. Our three datasets are then evaluated by humans through a survey. The process of creating the EmoMV dataset collection is visualized in Figure 1. A snapshot of some music video segments in this collection is depicted in Figure 2.

Figure 2: A snapshot of some music video segments in the EmoMV dataset collection.

In addition to the dataset creation, in this study, we also propose a deep neural network for classifying whether a pair of music and video is matched (i.e. carries similar emotional information) or mismatched. We train our proposed model in a multi-task learning manner, in which the binary classification (for the "matched" and "mismatched" labels) together with emotion classification on the video and music streams is carried out simultaneously. This model is then modified to tackle the affective music-video retrieval task. Our proposed model outperforms state-of-the-art approaches on the EmoMV dataset collection. Our experimental results provide a strong benchmark for affective correspondence learning between music and video modalities.

In summary, this study has made the following contributions:

(1) A collection of three new datasets for affective correspondence learning between music and video modalities is constructed.

(2) A deep neural network for binary affective music-video classification is proposed together with its adaptation to the affective music-video retrieval task. Extensive experiments are conducted on the newly created dataset collection, thus providing benchmark results.

(3) In the process of creating datasets, a multimodal music video emotion classification model is trained on the MVED dataset. The trained network is then applied to the other two datasets (including the Music Mood dataset of the AudioSet ontology and our self-collected dataset of soundtrack music video segments) to predict emotion labels corresponding to music video segments. Notably, the network outperforms many state-of-the-art approaches in [15, 19].

In the next section, we first provide an overview of the related studies. In Section 3, we represent the process of creating the EmoMV dataset collection. The user study on the accuracy of the "matched" and "mismatched" labels offered in the EmoMV dataset collection is described in Section 4. Next, our proposed models for binary affective music-video correspondence classification and affective music-video retrieval are represented in Sections 5 and 6, respectively, together with results on our datasets followed by the conclusion in Section 7.

2. Related work

Audio-visual correspondence learning has received considerable attention in recent years. In addition to the review of existing datasets, in this section, we focus on studies on affective correspondence between music and video modalities as well as related work on audio-visual correspondence learning. Before diving into datasets and models, we first provide a brief overview on different emotion representations.

2.1. Emotion representations

Many studies [21, 22, 23] have shown the influence of the multimedia such as music and videos on human emotions. Human emotions can be represented using continuous dimensions [24, 25] or discrete categories [26, 27]. In the continuous approach, emotions are mapped into a dimensional space [28, 29]. The Circumplex model proposed by Russell [28] is a continuous emotion representation model which is commonly used in many affective computing studies [30, 31, 32]. Other less popular continuous emotion representation models include the Vector model [33], and the Positive Activation - Negative Activation model [34, 35]. In the Circumplex model, there are two dimensions called arousal (the energy of the emotion) and valence (the negativity or positivity of the emotion) used to annotate emotions. There exists a third dimension called dominance, which is often omitted because it is hard to annotate [31]. In the categorical approach, emotions are described in terms such as happy, fearful, sad, etc., with a various number of emotional categories. For instance, there are six basic emotions in [36], and 27 categories in [21]. In the Geneva Emotional Music Scales (GEMS) model [37], the number of emotion terms is up to 45 (in GEMS-45), which is grouped into nine different categories.

Both continuous and discrete emotion representation approaches have been used to represent emotions evoked from music [38, 32, 39, 40] and videos [41, 30, 31, 21]. The advantage of the continuous approach is that it can fully model the diversity and complexity of human emotions. Its drawback, however, is that some emotions, such as nostalgia, may be difficult to distinguish and represent in continuous dimensions [42, 5]. This issue may be overcome by using a discrete representation of emotions. However, we are then faced with the problem of defining the taxonomy. The number of emotional categories could be very high, for example, 305 mood tags, as in allmusic [1] (accessed on 23/03/2022). Since the discrete representation of emotions is often easier for non-experts to comprehend, we make use of categories of emotion to construct matched and mismatched music-video pairs in this study.

2.2. Datasets for affective audio-visual correspondence learning

Most datasets [20, 43, 44, 45, 46, 47, 48], which are used in audio-visual correspondence learning tasks, often stem from other studies such as action recognition and sound classification. For *affective* audio-visual correspondence learning, there are only a few datasets, and they are also created by making use of other available ones. Some notable datasets are IMEMNet [4], IMAC [18], and two datasets introduced in [5]. The IMEMNet [4] and IMAC [18] datasets are created from music and images. To the best of our knowledge, the two datasets in [5] are the only ones dedicated to affective correspondence learning between music and videos. Music-video pairs are constructed through crowd-sourcing, whereby annotators are asked about how common the emotions in the two streams are. The first one consists of $3,000$ music-video pairs (one half consists of matched pairs and the other half are mismatched pairs in terms of emotions). In that dataset, the videos came from Cowen's dataset [21] and the music segments were randomly selected from the Unbalanced Train set of the Music Mood dataset of the AudioSet ontology [20]. In the second dataset in [5], music was drawn from Spotify [2], and videos were collected from Instagram [3] and the Moments in Time dataset [49]. However, these affective audio-visual correspondence learning datasets are not released, therefore, it is challenging to use them for benchmarking.

To overcome the data scarcity for affective audio-visual correspondence learning, we construct a collection of three datasets by self-collecting music videos from YouTube as well as making use of existing datasets originally created for

[1] `https://www.allmusic.com`

[2] `https://www.spotify.com/`

[3] `https://www.instagram.com/`

affective content analysis. While there are a wide number of video and music datasets for emotion prediction, to our knowledge, only a few music video datasets have emotion annotations. Some notable music video datasets are DEAP [17], MuVi [16], MVED [19], etc. Among them, the MVED dataset [19] is the largest one, whereby music video segments are labelled by humans. Therefore, we leverage the existing emotion annotations in this dataset to create matched and mismatched music-video pairs in our EmoMV-A dataset.

Most music video segments in the MVED dataset are from official music videos released by artists or producers. Therefore, to diversify our collection of datasets for affective audio-visual correspondence learning, in addition to the MVED dataset, we have also used real-world music videos (i.e. those that may be both raw or edited) to create matched and mismatched music-video pairs. Among the available music datasets with emotion annotations, the Music Mood dataset of the AudioSet ontology [20] is large. This dataset contains 10-second music segments obtained from clips on YouTube, whereby only the music stream is annotated by humans with mood labels (funny, happy, tender, sad, exciting, scary, and angry). The content of most music video segments in the Music Mood is generated by YouTube users, which makes it different from the MVED dataset. Therefore, we also make use of music video segments from the Music Mood dataset to create our own dataset (i.e. the EmoMV-B dataset). In addition to the use of available datasets, in this study, we also self-collect music videos of songs featured in movies that are available on YouTube and use them as the source to create the third dataset (i.e. EmoMV-C) in the EmoMV collection.

2.3. Affective audio-visual correspondence learning models

According to [2], a common approach to tackle the AVC task is to build a neural network, which includes three subnetworks: vision, audio and fusion. The vision and audio subnetworks are used to extract visual and audio features respectively, while the fusion subnetwork is used to combine those features to finally decide whether a pair of visual and audio modalities are matched or mismatched. Some pioneer three-subnetwork models are the L^3-Net [2] (for video frames and audio matching), and the Audio-Visual Embedding Network (AVE-Net) (for determining the location of a sound source in an image and cross-modal retrieval). The three-subnetwork approach [2] is also applied in various studies on affective correspondence learning between *music and images*, such as the affective correspondence prediction network (ACP-Net) in [18], and the model proposed in [50] that is similar to the ACP-Net.

For affective correspondence learning between *music and videos*, the music-video retrieval task is often the main focus (rather than just obtaining the "matched" and "mismatched" labels). Hence, instead of constructing a fusion subnetwork, a cross-modal distance learning subnetwork is often used, whereby the distance between the visual and audio embeddings is computed. In particular, in some notable studies [5, 51] on the emotion-aware music-video retrieval task, the proposed models consist of three subnetworks: vision, audio, and cross-modal distance learning. In [51], the authors propose a model called Acousticvisual Emotion Gaussians (AVEG) to learn the relationship among music, video, and emotion using the DEAP dataset [17]. The acoustic and visual features are first preprocessed using cross-modal factor analysis [52] before being passed to the AVEG model to obtain the predicted emotions corresponding to the music and videos. Music and video are then matched based on the similarity between the two predicted emotion distributions. In [5], in the vision subnetwork, the RGB stream of the pretrained Inflated-3D model (I3D) [53] is used as the visual feature extractor, while a CNN structure, which is similar to the network proposed in [54], is used in the audio subnetwork. The performance of the proposed model [5] is evaluated on three tasks: emotion classification for each modality, binary classification (on"matched" and "mismatched" labels) for music-video pairs, and cross-modal music retrieval.

In this study, inspired by the above affective audio-visual correspondence learning models, we propose a deep neural network, which includes three subnetworks (video, music and fusion), as a benchmark on the EmoMV dataset collection. In existing studies [5, 4, 18], the focus lies on spatial features that carry information about the appearance of objects appearing in videos, while information about the *motion* of objects is often ignored. The use of both spatial and temporal features carrying information about the appearance and motion of objects appearing in videos plays an essential role in video-related tasks, for instance, emotion prediction/classification [55, 56, 57, 58], and action recognition [59, 60]. Many pretrained networks such as two-stream CNN [59], FlowNet Simple (FlowNetS) [61], I3D [53], which were originally proposed for the action classification and detection task, can be used as feature extractors to obtain motion features of objects appearing in videos. In [62], the SlowFast network outperforms many state-of-the-art structures for action recognition on many datasets such as AVA [63], Charades [64], and Kinetics [65, 44]. The SlowFast structure consists of the Slow pathway (to capture the semantics from video frames), the Fast pathway (to obtain the motion), and lateral connections to fuse them. Therefore, in the video subnetwork of our proposed model, we apply the pretrained SlowFast network to obtain spatio-temporal features from the video stream.

To extract features from audio, we can use available tools (such as OpenSMILE [66] and YAFFE [67]), or pretrained deep neural networks (such as VGGish [68] and SoundNet [43]), which have been used for many tasks such as emotion prediction/recognition [56, 55, 69, 70, 71, 72], cross-modal audio-visual retrieval task in [73, 74]. In [72], the use of audio features extracted by VGGish (with parameters pretrained on the AudioSet dataset [20]) improves the performance of a speech emotion recognition model in comparison to using those obtained by the SoundNet network [43] and the OpenSMILE toolkit [66]. The VGGish network is also used in the cross-modal audio-visual retrieval task in [73, 74]. Therefore, in the music subnetwork of our proposed model, the VGGish structure pretrained on the AudioSet [20] dataset is used to extract audio features from the music stream. These extracted feature vectors are also released together with our newly created datasets.

To embed the audio and visual features into a common representation space, stacks of fully-connected layers are mainly applied in affective correspondence learning between music and videos [18, 5]. In this study, inspired by the projection heads (originally applied on the textual and visual features) in the Contrastive Language-Image Pretraining (CLIP) model [75], we construct music and video projection heads to embed the audio and visual features into a common representation space. In addition, to combine the audio and visual embeddings, there are many fusion techniques. Among them, the compact bilinear pooling [76] is commonly used to capture the complex associations between two modalities by enabling a multiplicative interaction between all the elements of two component vectors. This fusion technique is applied in many studies to combine, for example, visual and audio features for emotion recognition [77, 78], or textual and visual features for visual question answering [76]. Therefore, the multimodal compact bilinear pooling [76] is applied in this study to fuse visual and audio embeddings, resulting in our model to outperform many other approaches.

In addition to tackling the binary affective music-video correspondence classification task, we also modify our proposed model to adapt it to the affective music-video retrieval task. Particularly, the distance between the visual and audio embeddings is computed, instead of fusing the embeddings together to make a classification. Multi-task models have been introduced in affective computing, and have shown promising results. For instance, in [79], a multi-task attention network is proposed to predict emotions (represented in valence and arousal dimensions) and recognize facial expressions at the same time. In [80], a multi-task framework is also developed to simultaneously perform emotion regression and classification. In this study, we train our proposed model in a multi-task learning manner, whereby binary classification (for “matched” and “mismatched” labels) together with emotion classification (on the video and music streams) are simultaneously carried out. A similar approach is also applied to the retrieval task.

Table 1: Number of music video segments corresponding to each emotion label in the reduced MVED dataset (i.e. without the "neutral" label).

Reduced MVED dataset	Exciting	Fearful	Tense	Sad	Relaxing	**Total**
Train set	843	828	652	730	1,057	**4,110**
Validation set	102	111	84	111	148	**556**
Test set	50	50	50	50	50	**250**

3. Dataset creation

Due to the scarcity of publicly available large-scale datasets for affective correspondence learning between music and videos, we have created the EmoMV dataset collection. The process of creating each of the three datasets in this collection is described in detail below.

3.1. EmoMV-A dataset

In this study, the first dataset of the EmoMV collection (called EmoMV-A) is created by making use of music video segments from the MVED dataset [19]. The MVED dataset was originally created for the emotion classification task. To our knowledge, it is also the largest available dataset for this task with human-labelled emotions. The MVED dataset includes 5743 music video segments (with a duration of 30 seconds each) annotated with emotion labels (exciting, fearful, tense, sad, relaxing, and neutral). According to [19], the "neutral" label represents a mix of stimuli from the other five emotions, therefore, we exclude the segments annotated with this label. For convenience, we refer to the MVED dataset without the "neutral" music video segments as the *reduced MVED dataset*. This term will be used throughout this paper. After discarding the "neutral" music video segments, the number of music video segments remaining from the train, validation, and test sets of the reduced MVED dataset is 4110, 556, and 250 respectively. The number of music video segments corresponding to each emotion label is shown in Table 1.

"Matched"/ "Mismatched" labelling According to [19], in the MVED dataset, in the emotion annotation process, emotion labels are assigned to 30-second music video segments based on both visual and audio clues. For example, exciting music videos often contain smiling faces, dancing scenes, colouring effects, etc. in the video stream, and high pitch, large pitch variations, etc. in the music stream. In contrast, sad music video segments often contain slow-changing scenes, dark background, tears on faces, etc. in the video stream, with a slow tempo, soft music, etc. in the music. We refer readers to paper [19] for more detail on the emotion annotation process for the MVED dataset. Thanks to this carefully directed annotation process, we may consider the music and video streams from each music video segment taken from the MVED dataset as *matched* in terms of emotions. Therefore, the main challenge when creating the EmoMV-A dataset is how to create music-video pairs that are *mismatched* in terms of emotions. Our goal is to create a dataset consisting of music videos that are either matched or mismatched in terms of the emotions conveyed by their music and video streams. Therefore, given music video segments with emotions labelled on the music and video streams, the process of creating matched and mismatched music-video pairs is described as follows:

Step 1: We first divide each set (i.e. the train set, validation set, and test set) in the reduced MVED dataset into two clusters such that the number of music video segments corresponding to each emotion label appearing in one cluster is equal to that in the other. In particular, from the training set of the reduced MVED dataset as shown in Table 1), 843 exciting music video segments are divided into two subsets with 422 and 421 segments respectively; 828 fearful segments are divided into two subsets of 414 segments each, etc. The first cluster will form our matched samples, while the second cluster will be used to create our mismatched ones.

Step 2: The music video segments in the first cluster are kept intact, and are later used as *matched* music-video pairs. Those from the second cluster are used to create mismatched music-video pairs, in which a video stream assigned with this emotion label is paired with a music stream that is labelled with another

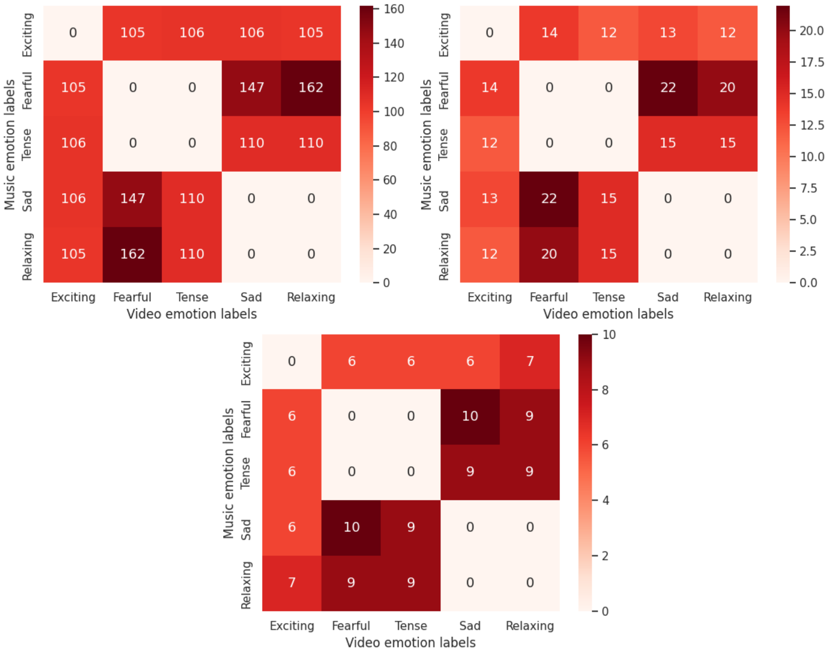


Figure 3: Number of the *mismatched* music-video pairs corresponding to each emotion label in the train (top left), validation (top right), and test (bottom) sets in the EmoMV-A dataset.

emotion. For doing so, we first create eight pairs of (mismatched) emotion labels: exciting – fearful, exciting – tense, exciting – sad, exciting - relaxing, fearful - sad, fearful – relaxing, tense – sad, and tense – relaxing. Note that according to [19], in the MVED dataset, the music video segments annotated with fearful or tense labels have different visual elements but possess the same audio components such as high pitch and high rhythmic variation. Similarly, those labelled with sad or relaxing emotions have common audio characteristics such as slow tempo and soft music. Therefore, the two emotion pairs (fearful – tense, and sad – relaxing) are not used.

Step 3: Based on the eight pairs of emotion labels, mismatched music-video pairs are constructed from the music video segments in the second cluster. In particular, an exciting video is paired with fearful music, and vice versa; an exciting video is paired with tense music, and vice versa, etc. The music video

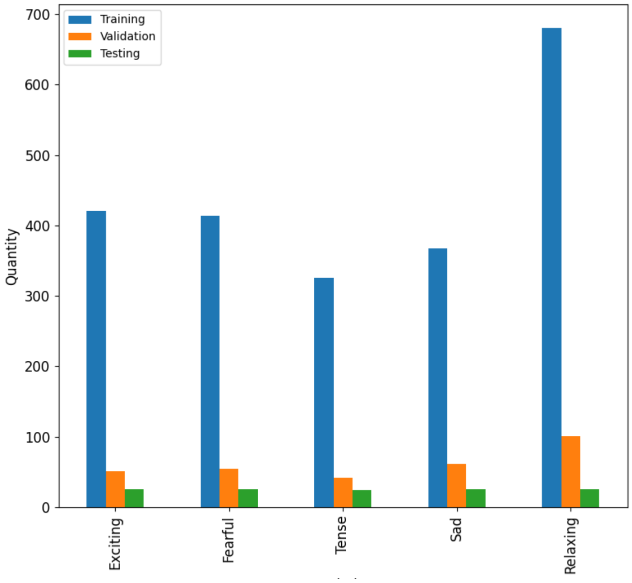


Figure 4: Visualization of the number of *matched* music video segments corresponding to each emotion label in the EmoMV-A dataset.

segments (belonging to the second cluster), which are redundant after this mismatching process, are originally the matched ones. Therefore, they are added to the first cluster.

As a result, our train set consists of $1,902$ mismatched music-video pairs, and $2,208$ matched ones. There are 310 matched and 246 mismatched pairs in the validation set. The number of matched and mismatched pairs in the test set is 126 and 124, respectively. However, to make the test set balanced with regard to the number of matched and mismatched pairs, we randomly exclude two matched music-video pairs (i.e. two music video segments) from this set. Hence, the final test set consists of 248 music-video pairs in total, whereby one half is matched, and the other half is mismatched. We visualize the number of mismatched music-video pairs with the corresponding emotion labels assigned to their video and music streams in Figure 3. The number of matched music-video pairs is plotted in Figure 4.

3.2. EmoMV-B dataset

To create the EmoMV-B dataset, we made use of the Music Mood dataset (of the AudioSet ontology), which consists of 10-music segments derived from music clips on YouTube. In the Music Mood dataset [4] (accessed on 16/10/2021), only the music stream is human-labelled with mood tags (funny, happy, tender, sad, exciting, scary, and angry), and many music segments are assigned multiple emotion tags. This dataset includes the Unbalanced Train set (consisting of an unbalanced number of music segments corresponding to each mood label), the Balanced Train set (consisting of an identical number of music segments corresponding to each mood label), and the Balanced Validation set, whereby only the music emotion annotations in the Balanced Train set (including 421 music segments) and the Balanced Validation set (consisting of 420 music segments) are verified by the authors. Those in the Unbalanced Train set are not verified, although this set is large with 1338 happy, 1035 funny, 1650 sad, 3971 tender, 5518 exciting, 985 angry, and 1617 scary segments. Most of the YouTube music clips, from which the Music Mood dataset is created, are user-generated. Hence, the use of this dataset diversifies our collection of datasets for the affective audio-visual correspondence task.

3.2.1. Data preprocessing

After downloading the music videos from YouTube, we manually validate every single music segment in the Music Mood dataset with the aim to filter out the footage of video games, as well as those that contain only some unrelated images in the video stream, etc. The music segments that are of low quality or mainly contain speech are also taken out. Due to these reasons, in addition to the unavailability of many given YouTube links, we obtain only 4487 music video segments in total.

[4] `https://research.google.com/audioset//ontology/music_mood_1.html`

3.2.2. Emotion labelling

In the Music Mood dataset, the emotion annotation process was carried out on the music stream only, while the video stream was not taken into account. Therefore, the emotion annotation provided in the Music Mood is not suitable for our goal. Emotion labelling requires a large number of annotators to reduce its subjectivity. Due to the high cost of the emotion annotation by humans, in this study, we apply a model (with parameters trained on another dataset) as an automatic tagging tool to assign emotion labels to the $4,487$ music video segments (filtered from the Music Mood dataset) as follows:

Emotion classification network: The Feature AttendAffectNet model [55] provides high accuracy in predicting the emotions of movie viewers represented in (continuous) valence and arousal values. According to [55], the model input can be either audio features or visual ones, or both; and the model performs best when both visual and audio features are used. Therefore, we first modify the Feature AttendAffectNet model by changing its last fully-connected layer from one neuron to six neurons followed by a softmax layer. This adaptation allows it to tackle the emotion classification task on the MVED dataset, whereby music video segments are annotated with one of six emotions (including exciting, fearful, tense, sad, relaxing, and neutral). For convenience, this model is called "modified Feature AttendAffectNet" throughout this paper. According to [55], the ResNet-50 [81], FlowNetS [61], and RGB-stream I3D networks [53] pretrained on the ImageNet dataset [82], the Flying Chairs dataset [61], and the Kinetics dataset [44], respectively, are used as feature extractors to obtain the visual features from the video stream. To obtain the audio features, the authors in [55] apply the OpenSMILE toolkit (with its "emobase2010" configuration, which is mainly for speech-related low-level feature extraction) and VGGish [68] pretrained on the AudioSet dataset [20]. The MVED dataset consists of music video segments rather than speech, therefore, in this study, we apply all of the aforementioned feature extractors, except for the OpenSMILE toolkit, to obtain visual and audio features from music video segments in the

MVED dataset.

Training details: We train, validate, and test the modified Feature AttendAffectNet structure on the MVED dataset. We consider three contexts, in which the model input either includes visual features only, audio features only, or both. In particular, the model is trained using the Adam optimizer, and the maximum number of epochs is 200. The batch size and the learning rate are 30 and 0.0005, respectively. The early stopping is applied with the patience parameter of 30. The cross-entropy loss is used in the training process, and the experiments are conducted in Python 3.6 using a NVIDIA GTX 1070 GPU.

Table 2: Performance of the modified Feature AttendAffectNet model on the test set of the MVED dataset.

Models	Accuracy(%)	F1-score	AUC
Feature AAN (video)	80.0	0.799	0.958
Feature AAN (music)	82.33	0.786	0.945
Feature AAN (music video)	86.67	0.866	0.982
Separable Slow-Fast network [15]	77.00	0.77	0.940
MVF (video, music, facial expression) [19]	74.00	0.73	0.926

Model performance on the MVED dataset: The performance of the modified Feature AttendAffectNet model on the MVED dataset is evaluated based on the classification accuracy, F1-score [83], the confusion matrix [84], and the Area Under the Receiver Operating Characteristics (AUC) [85]. As shown in Table 2, when using only visual features as the model input, the classification accuracy is 80.0%. The accuracy increases to 82.33% when using only audio features as the model input. When using both visual and audio features, the model reaches the highest classification accuracy of 86.67%. Notably, the modified Feature AttendAffectNet outperforms the state-of-the-art approaches including the Separable Slow–Fast network in [15] and the MVF model in [19] (as shown in Table 2), even when only visual features are used as its input. In addition to the prediction accuracy, we also visualize the confusion matrix in Figure 6 and the area under the ROC as shown in Figure 5.

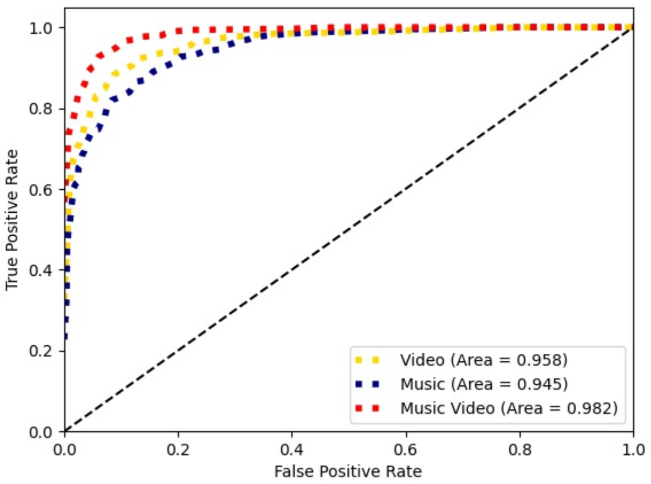


Figure 5: The ROC curve for emotion classification on the MVED dataset using the modified Feature AttendAffectNet with video only, music only, and both as the model input.

Table 3: Number of music video segments (from the filtered Music Mood dataset) that correspond to emotion labels predicted by the modified Feature AttendAffectNet.

Exciting	Fearful	Tense	Sad	Relaxing	Neutral	**Total**
225	56	72	268	126	85	**832**

Automatic emotion tagging To obtain emotion labels for music video segments in the filtered Music Mood dataset, we first extract visual and audio features from the video and music streams in each segment (by using feature extractors consisting of the pretrained ResNet-50, FlowNetS, RGB-stream I3D, and pretrained VGGish networks as mentioned above). We then feed only visual features, only audio features, and both of them to the modified Feature AttendAffectNet (which is modified for three input contexts: video only, audio only, and both, respectively) with parameters previously trained on the MVED dataset. Only music video segments for which the same emotion label is predicted for their video stream, music stream, and both, are chosen. As a result,

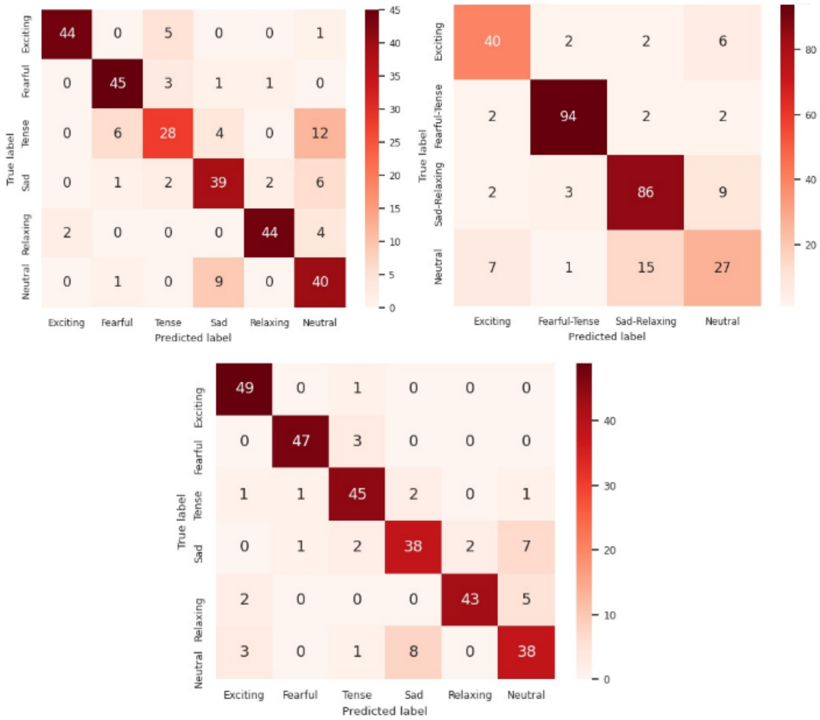


Figure 6: The confusion matrix when the modified Feature AttendAffectNet is applied to classify emotions for videos (top left), music (top right), and music videos (bottom).

we obtain a set of 832 music video segments as shown in Table 3.

3.2.3. EmoMV-B dataset creation

From the set of 832 music video segments obtained from the above automatic emotion tagging process as shown in Table 3, we discard those with the "neutral" label. As a result, we obtain 747 music video segments in total (225 exciting, 56 fearful, 72 tense, 268 sad, and 126 relaxing). We first divide these music video segments into two sets with a proportion of 80% and 20%, respectively. These sets are then used to create the train and validation sets, respectively.

Similar to the creation of the EmoMV-A dataset, each of these sets is then divided into two clusters, such that the number of music video segments corresponding to each emotion label appearing in one cluster is equal to that in

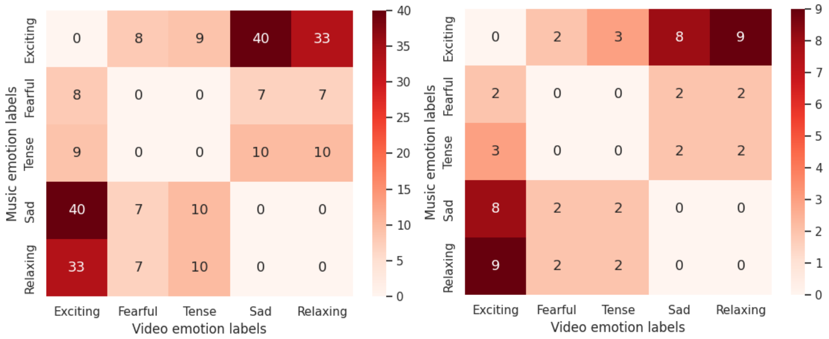


Figure 7: Number of *mismatched* music-video pairs with the corresponding emotion labels assigned to their music and video streams in the training (left), and validation (right) set in the EmoMV-B dataset.

the other. One cluster is used to create mismatched music-video pairs, while the other is kept intact. The process of creating mismatched music-video pairs is the same as that described in Subsection 3.1. As a result, the number of mismatched music-video pairs obtained in the train and validation sets is 248 and 60, respectively. The number of the mismatched music-video pairs with the emotion labels predicted for their video and music streams is visualized in Figure 7. With the aim to create a dataset that is balanced in the number of matched and mismatched music-video pairs, 248 and 60 music video segments from other clusters (that were originally kept intact) are used as the matched music-video pairs for the training and validation sets, respectively. The number of matched music-video pairs corresponding to each emotion label in the train and validation sets of the EmoMV-B dataset is illustrated in Figure 8. In a nutshell, the EmoMV-B dataset consists of 616 music video segments (496 for the train set and 120 for the validation set), and these sets are balanced in the number of matched and mismatched pairs.

3.3. EmoMV-C dataset

This dataset is created from the self-collected music videos of songs featured in movies (also called soundtrack music videos). Hence, many of them contain

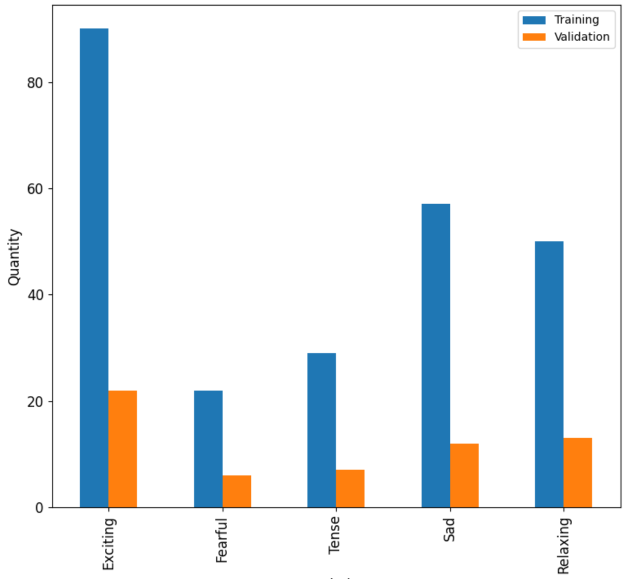


Figure 8: Visualization of the number of *matched* music-video pairs corresponding to each emotion label in the train and validation sets in the EmoMV-B dataset.

some movie scenes, which makes the EmoMV-C dataset different from the two previous ones. The creation process of this dataset is explained below.

3.3.1. Data collection

We first collect movie titles by using Google with different search terms, for example, "best movies", "happy movies", "sad movies", etc. We then search the titles of songs featured in the movies. These song titles are then used as keywords to find the corresponding music videos on YouTube. As a result, we collect 235 soundtrack music videos. In addition, we also collect another 119 soundtrack music videos based on the song titles given in [86]. The collected 354 soundtrack music videos are split into 2688 music video segments with a duration of 30 seconds each.

3.3.2. Emotion labelling

Similar to the creation of the EmoMV-B dataset, we apply feature extractors (i.e. the pretrained ResNet-50, FlowNetS, RGB-stream I3D, and VGGish networks) on the 2688 music video segments in order to obtain visual and audio features from their video and music streams. The extracted feature vectors are then passed to the modified Feature AttendAffectNet model (pretrained on the MVED dataset for emotion classification). We select the music video segments that have the same predicted emotion label regardless of whether the visual features only, or audio features only, or both are used as the model input. As a result, we obtain 663 music video segments (consisting of 189 music video segments labelled as “exciting”, 42 “fearful”, 69 “tense”, 148 “sad”, 198 “relaxing”, and 17 “neutral”) that satisfy this condition. These 663 music video segments originate from 336 music videos . Among these segments, many come from the same original music videos, therefore, they might contain repeated or similar content. To avoid this, we use no more than two segments from the same original music video and remove the redundant ones. As a result, after applying such filtering criteria, we obtain 570 music video segments (including 167 “exciting”, 35 “fearful”, 53 “tense”, 127 “sad”, 174 “relaxing”, and 14 “neutral”) as shown in Table 4.

Table 4: Number of music video segments (originating from soundtrack music videos) obtained after applying filtering criteria.

Exciting	Fearful	Tense	Sad	Relaxing	Neutral	**Total**
167	35	53	127	174	14	**570**

3.3.3. EmoMV-C dataset creation

In this step, after discarding the segments with the “neutral” label, we obtain 556 music video segments (originated from 325 soundtrack music videos) in total. These segments are divided into two sets, with proportions of 80% and 20%, which are later used to create the train and validation sets, respectively.

Note that we ensured that the music video segments belonging to the same original music videos appear in one of these sets only. By doing so, we prevent information from being leaked from the train set into the validation set.

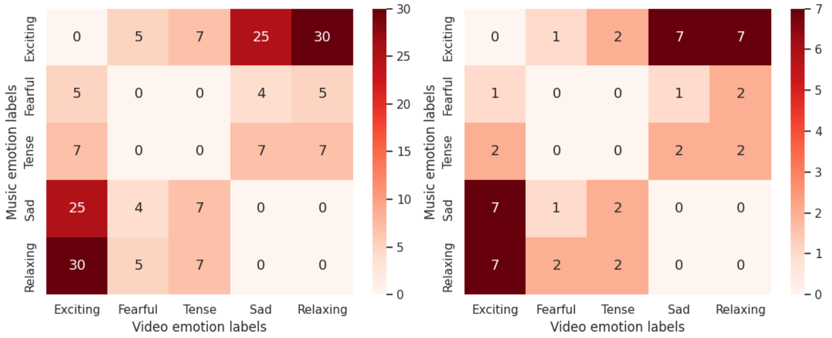


Figure 9: Number of *mismatched* music-video pairs corresponding to each emotion label in the training (left), and validation (right) sets in the EmoMV-C dataset.

We then divide each set into two clusters such that the number of music video segments corresponding to each emotion label appearing in each cluster is equal to that of the other. Similar to the process of creating the two previous datasets, one cluster is used to create the mismatched music-video pairs, while the other half is kept intact. In this dataset, the process of creating mismatched music-video pairs is the same as the one described in Subsection 3.1. As a result, we obtain 180 and 48 mismatched music-video pairs in the train and validation sets, respectively. A matrix of the number of mismatched music-video pairs together with the corresponding emotion labels assigned to their video and music streams is visualized in Figure 9. Similar to the EmoMV-B dataset, with the aim to create a dataset that is balanced in the number of matched and mismatched music-video pairs, 180 and 48 music video segments from other clusters (that are originally kept intact) are used to form the matched music-video pairs for the training and validation sets, respectively. We also illustrate the number of matched video-music pairs corresponding to each emotion label in the training and validation sets of the EmoMV-C dataset in Figure 10. In short, the EmoMV-C dataset consists of 456 music video segments, with 360

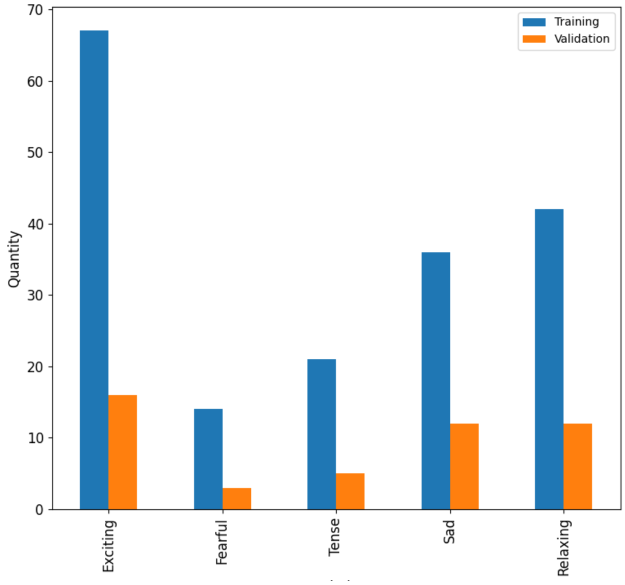


Figure 10: Visualization of the number of *matched* music video segments corresponding to each emotion label in the EmoMV-C dataset.

segments used for training and 96 segments for validation. Each set is balanced in terms of the number of matched and mismatched pairs.

4. Dataset evaluation

We conduct an online survey to evaluate the accuracy of the labels in our datasets, whereby the "matched" and "mismatched" labels as well as the emotion labels assigned to the music and video modalities are considered. The dataset verification process is described below.

4.1. Survey design

From each dataset in the EmoMV collection, we randomly select ten samples (i.e. music video segments) including five "matched" and five "mismatched". As a result, 30 samples in total are used in this survey. Note that in our datasets, the music and video streams in each music video segment are labelled with one of five emotions (exciting, fear, tense, sad, and relaxing). The proportion of samples (among the 10 samples selected from each dataset) corresponding to each of these emotion labels is identical.

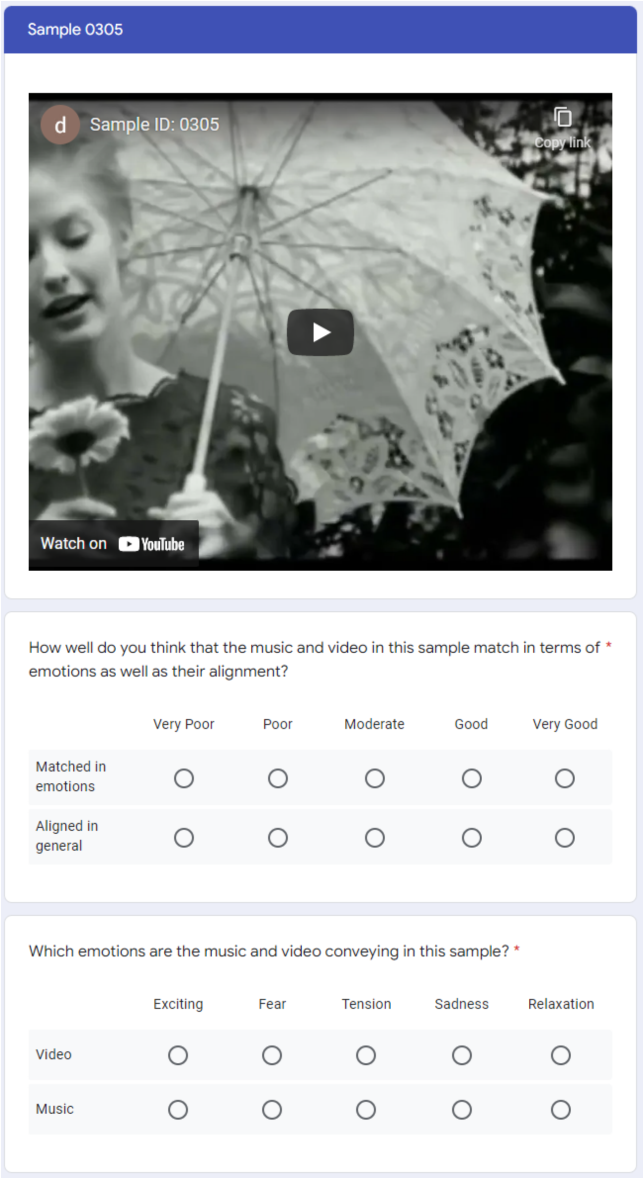
In our survey, participants are asked to listen to and view music video segments, and rate how well the music and video streams in each sample match in terms of emotions and their general alignment. They can indicate their answers on a 5-point Likert scale ("very poor", "poor", "moderate", "good", and "very good"). Participants are also asked which emotions (among the given five emotions) are conveyed in the music and video streams of each sample. Each participant is asked to watch all 30 samples. A snapshot of a sample with questions used in our survey is visualized in Figure 11.

4.2. Survey results

A total of 22 subjects participated in our survey. As a result, we obtained 660 ratings across all 30 samples (i.e. 220 ratings for ten samples from each dataset).

4.2.1. EmoMV-A dataset

As shown in Table 5, for the matched samples selected from this dataset, the proportion of "good - very good", "moderate", and "poor - very poor" ratings for the question related to the level of matching in terms of emotions is 77.27%, 14.55%, and 8.18%, respectively. For the question about the general alignment between music and video streams, this proportion is 79.09%, 12.73%, and 8.18%, respectively. For the mismatched samples selected from the EmoMV-A dataset, the proportion of "good - very good", "moderate", and "poor - very poor" responses for the emotion matching level is 6.36%, 10.91%, and 82.73%, respectively. These values for the general alignment between music and video streams are 9.09%, 12.73%, and 78.18%, respectively. This confirms that the matched segments are perceived as more matched, and the mismatched segments are mostly perceived as mismatched.



Sample 0305

How well do you think that the music and video in this sample match in terms of emotions as well as their alignment? *

	Very Poor	Poor	Moderate	Good	Very Good
Matched in emotions	○	○	○	○	○
Aligned in general	○	○	○	○	○

Which emotions are the music and video conveying in this sample? *

	Exciting	Fear	Tension	Sadness	Relaxation
Video	○	○	○	○	○
Music	○	○	○	○	○

Figure 11: A snapshot of a sample with questions used in our online survey.

Participants also indicated the emotions they thought are conveyed in the music stream. The proportion of ratings on the music stream for which the emotion label (originally from the MVED dataset [19]) is the same as the one obtained in our survey is 57.27%. For the video stream, this proportion is 64.55%. In general, these overlapping values are not high, even though both annotation processes are carried out by humans. This might be due to the fact that the emotion is subjective.

Table 5: Proportion of ratings for samples selected from the EmoMV-A dataset.

EmoMV-A dataset		**Ratings (%)**		
		Very poor - Poor	Moderate	Good - Very good
Matched samples	Matched in emotions	8.18	14.55	77.27
	General alignment	8.18	12.73	79.09
Mismatched samples	Matched in emotions	82.73	10.91	6.36
	General alignment	78.18	12.73	9.09

4.2.2. EmoMV-B Dataset

The proportion of “good - very good”, “moderate”, and “poor - very poor” ratings with regards to the matching level between the emotions conveyed in the music and video streams, is 78.18%, 10.91%, and 10.91%, respectively, in the EmoMV-B dataset. These values with regards to the general alignment between music and video streams are 80.91%, 12.73%, and 6.36%, respectively, as shown in Table 6. For the mismatched samples selected from this dataset, the proportion of “good - very good”, “moderate”, and “poor - very poor” responses with regards to the emotion-matching level is 25.45%, 12.73%, and 61.82%, respectively. These values for the general alignment between the two streams are 23.64%, 15.45%, and 60.91%, respectively. This rate is slightly below the one obtained on the EmoMV-A dataset, and we suspect this is because the matched and mismatched music-video pairs were constructed based on emotion labels predicted from a model versus those annotated by humans.

Looking at the emotion tags provided by the participants for the music stream, the proportion of emotion labels that are the same as the ones offered in our dataset is 77.27%, while this proportion for the video stream is 47.27%. This difference may be explained by the construction of the EmoMV-B dataset using music video segments from the Music Mood dataset. According to [14], music is effective in evoking emotions in viewers. Therefore, the emotions conveyed in the music stream might be more vivid than the ones carried in the video stream.

Table 6: Proportion of ratings for samples selected from the EmoMV-B dataset.

EmoMV-B dataset		**Ratings (%)**		
		Very poor - Poor	Moderate	Good - Very good
Matched samples	Matched in emotions	10.91	10.91	78.18
	General alignment	6.36	12.73	80.91
Mismatched samples	Matched in emotions	61.82	12.73	25.45
	General alignment	60.91	15.45	23.64

4.2.3. EmoMV-C Dataset

For the matched samples selected from the EmoMV-C dataset, the proportion of “good - very good”, “moderate”, and “poor - very poor” ratings with regards to the emotion-matching level is 91.82%, 4.54%, and 3.64%, respectively. The rate with regards to the general alignment between the two streams is 88.18%, 8.18%, and 3.64%, respectively, as shown in Table 7. For the mismatched samples selected from this dataset, the proportion of “good - very good”, “moderate”, and “poor - very poor” responses on the emotion-matching level is 11.82%, 13.64%, and 74.54%, respectively. This proportion with regards to the general alignment between music and video streams are 10.91%, 7.27%, and 81.82%, respectively. These rates show that the matched and mismatched samples are mostly differentiated by humans.

In addition, for the music stream, the overlapping rate between the labels provided by the participants in our survey and the ones offered in our dataset is 86.36%. For the video stream, this rate is 53.18%. Similar to the EmoMV-B dataset, this difference might be because the music is more effective in conveying emotions to listeners.

Table 7: Proportion of ratings for samples selected from the EmoMV-C dataset.

EmoMV-C dataset		**Ratings (%)**		
		Very poor - Poor	Moderate	Good - Very good
Matched samples	Matched in emotions	3.64	4.54	91.82
	General alignment	3.64	8.18	88.18
Mismatched samples	Matched in emotions	74.54	13.64	11.82
	General alignment	81.82	7.27	10.91

5. Binary affective music-video correspondence classification

5.1. Proposed model

Due to the limit in the number of available datasets together with benchmarks on affective audio-visual correspondence learning, in addition to creating three new datasets, we propose a deep neural network to classify whether music-video pairs are matched or mismatched in terms of emotions. The three-subnetwork approach has been proved to be effective in many audio-visual correspondence learning tasks. Following this approach, we can leverage pretrained deep neural networks (that are originally developed for action recognition and audio classification) to extract visual and audio features from video and music streams, respectively. Video and music projection heads are then applied to embed the visual and audio features into a common representation space. The obtained visual and audio embeddings are then fused, and a binary classification task is performed to predict "Yes" (i.e. "matched") or "No" (i.e. "mismatched") as illustrated in Figure 12. In addition, following the multi-task learning approach, we append the music and video branches (for music and video emotion classification) to the music and video subnetworks, respectively. Components of our proposed model are described in detail below.

Video subnetwork In the video subnetwork, we make use of the SlowFast network [62] pretrained on the Kinetics human action video dataset [44] (except for its last classification layer) to extract the spatial and temporal features

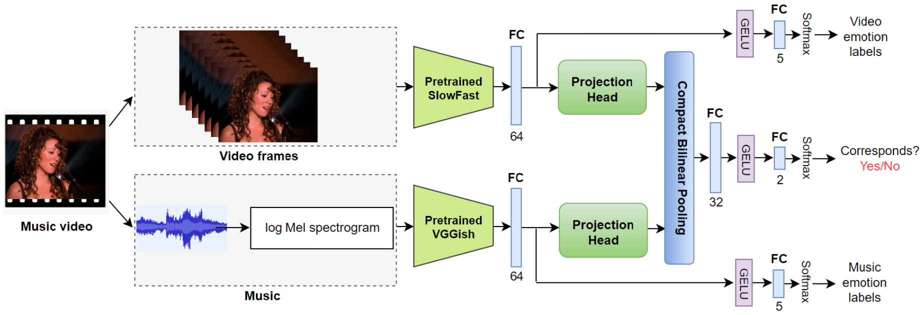


Figure 12: Our proposed model. The extracted visual and audio features are passed to fully-connected layers for dimensionality reduction before being projected into a common representation space. The visual and audio embeddings are then fused before being fed into other fully-connected layers, from which a binary classification task is carried out.

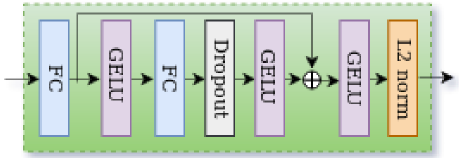


Figure 13: Projection head used to project visual and audio features into a common representation space.

that carry both appearance and motion information about objects appearing in the video stream. As a result, we obtain a $2,304$-dimensional feature vector from each music video segment. This feature vector is then fed to a fully-connected layer of 64 neurons for dimensionality reduction (as visualized in Figure 12) before being passed to the video projection head. This projection head consists of fully-connected layers of 64 neurons each, the Gaussian error linear unit (GELU) [87], a dropout ratio of 0.5, a residual connection, and $L2$-normalization, as described in Figure 13.

Music subnetwork In the music subnetwork, we apply the VGGish network [68] pretrained on the AudioSet ontology [20] to extract a 128-dimensional feature vector from each 0.98-second music segment (which is at a sampling rate of 16 kHz with signed 16-bit PCM encoding and a mono channel). We then compute the element-wise average of the obtained feature vectors extracted from all 0.98-second music segments from each music video segment. As a result, we obtain a 128− dimensional vector representing the music stream of each music video segment. Similar to the video subnetwork, this feature vector is then passed to a fully-connected layer of 64 neurons for dimensionality reduction. A music projection head, which has the same structure as the one for the video stream (as described in Figure 13), is also applied to embed the dimension-reduced audio feature vector into the common representation space.

Fusion subnetwork The multimodal compact bilinear pooling [76] is applied to the visual and audio embeddings (achieved by applying the video and music projection heads) to obtain their joint representation. The joint representation is then passed to a fully-connected layer of 32 neurons followed by the GELU activation function, a fully-connected layer of two neurons, and a softmax layer to obtain the output. In practice, we tried different fusion techniques such as concatenation and average pooling. We also adapt the fusion subnetwork proposed in the L^3-Net model in [2] to our model. However, these techniques do not perform better than the multimodal compact bilinear pooling [76] on the EmoMV dataset collection.

Emotion classification branches We perform multi-task learning by appending video and music branches to the video and music subnetworks. In particular, after being fed to the fully-connected layers of 64 neurons in the video and music subnetworks for dimensionality reduction, the dimension-reduced visual and audio feature vectors are then passed to the newly added video and music emotion classification branches, respectively. These branches have the same structure, which consists of a GELU [87] followed by a fully-connected layer of five neurons and a softmax layer to obtain the emotion classification output of each respective modality. We jointly train these emotion classification branches with the binary classifier, whereby three cross-entropy loss functions are used, and each of them is equally weighted.

5.2. Baseline model

Due to the scarcity of benchmarks for affective video-music correspondence learning, to assess the performance of our proposed model, we modify the model proposed in [5] (originally developed for the music retrieval given videos as queries such that they convey a similar emotion) and use it as a baseline. The model in [5] includes a video subnetwork, a music subnetwork, and a cross-modal learning distance subnetwork. The video subnetwork consists of the RGB stream of the I3D network [53] pretrained on the Kinetics dataset [44] (which is used a feature extractor) followed by fully connected layers. Each fully-connected layer is followed by a rectified linear unit (ReLU), except for the output layer, which is followed by a sigmoid function. The music subnetwork is similar to the network proposed in [54], which includes a block of convolutional layers, each is followed by the batch normalization and a ReLU activation function. In the music subnetwork, the output layer is a fully-connected layer followed by a sigmoid function. In the cross-modal distance learning subnetwork of the model in [5]], the visual and audio feature vectors are projected into a common representation space (by using two fully-connected layers followed by the L2 normalization), from which their Euclidean distance is computed. According to [5], the authors train the model with binary cross-entropy loss (for emotion classification in video and music subnetworks), and contrastive loss [88] (on the L2-normalization of the obtained Euclidean distance in the cross-modal distance learning subnetwork). Although the model in [5] is designed for the music-video retrieval task, it is also used as a binary classifier (by passing the obtained L2-normalized Euclidean distance to a sigmoid function) with a threshold of 0.5 to classify whether music-video pairs are matched or mismatched in terms of emotions. We refer readers to Figure 2 in the paper [5] for more detail. For a fair comparison with our proposed model, the model [5] is implemented with the same visual and audio features as those used in ours (i.e. the feature vectors extracted by using the pretrained SlowFast and VGGish networks as mentioned in Section 5.1). We train this model in a multi-task learning manner. However, it does not perform well on the EmoMV dataset collection (with the classification accuracy of less than 50%). To improve the performance of the model, we train it in two stages. In the first stage, we perform emotion classification on the music and video streams in a multi-task learning manner, in which two cross-entropy loss functions together with the contrastive loss (on the L2-normalization of the Euclidean distance between the visual and audio embeddings). Note that in this stage, we do not apply the sigmoid function (with a threshold of 0.5 to classify whether music-video pairs are matched or mismatched in terms of emotions) as conducted in [5]. In the second stage, a logistic regression model is trained on the L2-normalization of the obtained Euclidean distance between the visual and audio embeddings to predict matched and mismatched music-video pairs. We use this framework as a baseline model in this study.

5.3. Experiments

The performance of our proposed model is evaluated on the EmoMV collection. The experimental setup and the obtained results are described below.

5.3.1. Experimental Setup

In this study, the Adam optimizer is used in the training phase, whereby the maximum number of epochs is 1000. We set the batch size and the learning rate to 256, and 0.0001, respectively. The early stopping is applied with the patience parameter of 20. We conduct the experiments using Python 3.6 on a NVIDIA GTX 1070 GPU. This setup is applied for both our proposed model and the baseline.

5.3.2. Results

We carry out experiments on three datasets of the EmoMV collection. In addition to the classification accuracy, we also use the F1-score [83], and the AUC score [85] to evaluate the performance of our proposed model.

As shown in Table 8, our proposed approach performs better than the baseline model on all three datasets of the EmoMV collection. Using our model, the classification accuracy, the F1-score, and the AUC obtained on EmoMV-A are 79.03%, 0.80 and 0.87, respectively. These values are much higher than those

Table 8: Performance of our proposed model on the EmoMV dataset collection with multi-task learning

Dataset	Multi-task learning models	Accuracy(%)	F1-score	AUC
EmoMV-A	Baseline (Modified model [5])	75.00	0.75	0.82
	Our model	**79.03**	**0.80**	**0.87**
EmoMV-B	Baseline (Modified model [5])	65.00	0.64	0.75
	Our model	**75.00**	**0.76**	**0.81**
EmoMV-C	Baseline (Modified model [5])	67.71	0.65	0.73
	Our model	**70.83**	**0.68**	**0.74**

obtained by using the baseline model (with the accuracy, the F1-score, and the AUC of 75.00%, 0.75, and 0.82, respectively). On the EmoMV-B dataset, using our model, these values are 75%, 0.76, and 0.81, respectively. Our model performs worst on the EmoMV-C dataset with the classification accuracy of 70.83%, the F1-score of 0.68, and the AUC of 0.74. However, its performance is still better than the baseline with the accuracy of 67.71%, the F1-score of 0.65, and the AUC of 0.73 on this dataset. In general, our proposed model as well as the baseline perform on the EmoMV-A dataset much better than on the EmoMV-B and EmoMV-C datasets. This could be due to the fact that the matched and mismatched music-video pairs in the EmoMV-A dataset are constructed from music video segments with human-annotated emotion labels, while those used to create the EmoMV-B and EmoMV-C datasets are automatically tagged by applying the modified Feature AttendAffectNet model. Therefore, these datasets might contain some music-video pairs with noisy labels, and this might be difficult for the models to classify whether music-video pairs are matched or mismatched.

We also compute the emotion classification accuracy on music and video streams of our proposed model and the baseline on the EmoMV dataset collection. On the EmoMV-A dataset, the emotion classification accuracy of our model is 65.32% on music and 47.18% on video stream. These values are 83.87% and 39.92%, respective when the baseline model is used. On the EmoMV-B dataset, our model achieves 41.67% accuracy on music and 47.50% on video stream. These values are 65.00% and 45.00% when the baseline model is used. On the EmoMV-C dataset, the accuracy of the baseline model is 58.33% for music and 43.75% for video stream, while our model achieves 50.00% accuracy for music and 40.63% for video stream. The baseline approach outperforms our model on the emotion classification task on music and video streams. This might be because the baseline model, in nature, is designed for emotion classification (as mentioned in Subsection 5.2). In addition, when developing our model, the binary affective correspondence classification task is our main focus. In general, both the baseline and our approach do not perform well on the emotion classification task (on music and video streams) on the EmoMV dataset collection. This phenomenon can be explained by the fact that learning multiple tasks simultaneously is a challenging optimization problem, which might result in lower overall performance in some cases in comparison to learning each task separately as mentioned in [89, 90].

5.3.3. Ablation Study

Table 9: Performance of our model on the EmoMV dataset collection with single task learning.

Dataset	Single task learning models	Accuracy(%)	F1-score	AUC
EmoMV-A	Baseline (Modified model [5])	62.10	0.62	0.70
	Our model	**80.65**	**0.80**	**0.87**
EmoMV-B	Baseline (Modified model [5])	58.33	0.57	0.53
	Our model	**75.83**	**0.76**	**0.85**
EmoMV-C	Baseline(Modified model [5])	59.38	0.57	0.61
	Our model	**67.71**	**0.66**	**0.77**

Instead of doing multi-task learning, we remove the emotion classification branches from our proposed model as well as the baseline. These reduced models are trained using the experimental setup described in Subsection 5.3.1. As shown in Tables 8 and 9, on all three datasets of the EmoMV collection, there is no considerable difference in our model performance when single task learning or multi-task learning is carried out. In addition, whether the music and video branches are appended to the models or not, our approach also achieves a better performance than the baseline on the EmoMV collection. In particular, after removing the music and video branches, the accuracy of our model is 80.65% on the EmoMV-A dataset. This value is 75.83% and 67.71% on the EmoMV-B and EmoMV-C datasets, respectively. The accuracy, together with the F1-score, and the AUC of the baseline model on the EmoMV-A dataset is 62.10%, 0.62, and 0.70, respectively, after the music and video branches are removed from its structure. The accuracy of the baseline model on the EmoMV-B and EmoMV-C datasets is 58.33%, and 59.38%, respectively. The F1-score and the AUC of the baseline model on the EmoMV-B dataset are 0.57 and 0.53, respectively. On the EmoMV-C dataset, these values are 0.57, and 0.61, respectively.

6. Affective Music-Video Retrieval

In this study, in addition to tackling the binary affective music-video correspondence classification task, we also do affective music-video retrieval. Particularly, given a (muted) video segment as a query, our model retrieves relevant music segments (i.e. music segments conveying a similar emotion as the video query), and vice versa. The adaptation of the above proposed model to the affective music-video retrieval task, together with evaluation metrics and the model performance on the EmoMV dataset collection are described in detail below.

6.1. Model Adaptation to Affective Music-Video Retrieval

Our proposed model (as shown in Figure 12) is originally designed for the binary classification task on “matched” or “mismatched” labels. To adapt it to affective music-video retrieval, we compute the distance between visual and audio embeddings, instead fusing them together. This approach is similar to the ones proposed in [4, 5, 51]. In particular, the compact bilinear pooling together with the following fully-connected layers and the softmax function are discarded. Instead, we compute the cosine distance $d_{cos}(f_v, f_m)$ between the visual and audio embeddings (denoted as f_v, and f_m, respectively) as follows:

$$d_{cos}(f_v, f_m) = 1 - S_{cos}(f_v, f_m), \tag{1}$$

where $S_{cos}(f_v, f_m)$ is the cosine similarity between the visual embedding (f_v) and audio embedding (f_m). The cosine similarity is computed by using the following formula:

$$S_{cos}(f_v, f_m) = \frac{f_v . f_m}{\|f_v\| \times \|f_m\|}, \tag{2}$$

where $\|f_v\|$ and $\|f_m\|$ are the Euclidean norm of vectors f_v and f_m, respectively.

The music-video retrieval network is trained jointly with the video and music emotion classification branches using the same experimental setup mentioned in Subsection 5.3.1, except for the loss functions. In particular, we use three equal-weighted loss functions including two cross-entropy loss functions (for the video and music subnetworks), and the contrastive loss [88] on the cosine distance between the visual and audio embeddings. We train our model as well as the baseline in a multi-task learning manner. In the inference process, given a video as a query, our model computes the cosine similarity score (as described in Equation 2) between the visual embedding (of the video query) and the audio embeddings of all given music segments (in the database). Based on this similarity score, all given music segments are ranked, and the top-ranked results are considered to be the best matches to the video query. For the baseline model, the similarity score S_{Euclid} between the visual and audio embeddings is computed from their Euclidean distance d_{Euclid} as follows:

$$S_{Euclid} = \frac{1}{d_{Euclid}}. \tag{3}$$

6.2. Experiments

We evaluate the performance of our proposed model for affective music-video retrieval on the EmoMV dataset collection. The evaluation metrics together with the obtained results are described below.

6.2.1. Evaluation Metrics

To evaluate the performance of our model on the affective music-video retrieval task, we compute the Mean Average Precision score as used in [91, 92, 93], and the top-K retrieval accuracy. These scores are defined as follows:

Top-K retrieval accuracy Given videos as queries, the top-K retrieval accuracy is the proportion of the video queries with at least one relevant music segment retrieved within the first K results. In [5], only top-1 retrieval accuracy is used.

Mean Average Precision (mAP) According to [92, 93], when computing the mAP score, a retrieved result is considered as *relevant* to the query if it has the same label as the query, otherwise, it is irrelevant. This means that for each video, there are many relevant (i.e. ground-truth) music segments and vice versa. In the information retrieval theory [94], the mAP is the mean of the Average Precision (AP) of all queries, and is defined as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{4}$$

where N is the number of queries, AP_i is the Average Precision for query i, which is computed using the following formula:

$$AP_i = \frac{1}{R_i}\sum_{j=1}^{R_i} Precision_i(rel = j), \tag{5}$$

where R_i is number of relevant documents for query i (Note that in our study, the documents can be understood as music segments if the query is a video, and vice versa). $Precision_i(rel = j)$ is the precision at the j-th document that is *relevant* to query i and is computed as follows:

$$Precision_i(rel = j) = r_i(j)/j, \tag{6}$$

where $r_i(j)$ is the number of documents up to position j that are relevant to query i.

6.2.2. Results

We use the test set of the MVED dataset, from which the test set in our EmoMV-A dataset is constructed, to evaluate the performance of our music-video retrieval model. Note that our affective music-video retrieval model is trained on the train set of the EmoMV-A dataset, in which the “matched” and “mismatched” music-video pairs are constructed from music video segments annotated with “exciting”, “tense”, “fearful”, “sad”, and “relaxing” labels (except for the “neutral” ones). Therefore, we discard the “neutral” music video segments from the test set of the MVED dataset before using it to evaluate the performance of our music-video retrieval model. Similarly, the music video segments (except for the “neutral” ones), from which the validation sets of the EmoMV-B and EmoMV-C datasets are constructed, are also used to evaluate the performance of our retrieval model.

As shown in Tables 10, and 11, our model outperforms the baseline for the affective music-video retrieval task on all three datasets of the EmoMV collection. Using videos as queries, on the EmoMV-A dataset, the top-1 accuracy of our model is 56.00%, while that of the baseline is 34.40%. Similarly, our model also reaches higher top-3 and top-5 accuracy on this dataset. Our mAP is 58.53%, whereas the baseline achieves 39.59% only for this score on the EmoMV-A dataset. On the EmoMV-B and EmoMV-C datasets, the top-K accuracy, as well as the mAP of our model, are not as high as those on the EmoMV-A dataset, however, they are still better than those obtained by using the baseline model. On the EmoMV-B dataset, the baseline model reaches 33.33% for the top-1 accuracy, which is much lower than 43.33% achieved by using our model. In terms of the mAP, our model achieves 46.14%, yet the baseline only reaches 40.79%. On the EmoMV-C dataset, our top-1 accuracy and mAP are 33.33%, and 41.64%, respectively. These values are 27.08%, and 38.82%, respectively when the baseline model is used.

When music segments are used as queries to retrieve videos, the top-1 accuracy and the mAP of our model on the EmoMV-A dataset are 56.00% and

Table 10: Given a video query, retrieve music: Accuracy and the mAP on the EmoMV dataset collection.

Dataset	Multi-task learning models	Top-1(%)	Top-3(%)	Top-5(%)	mAP (%)
EmoMV-A	Baseline (Modified model [5])	34.40	60.80	68.00	39.59
	Our model	**56.00**	**70.80**	**75.20**	**58.53**
EmoMV-B	Baseline (Modified model [5])	33.33	45.83	66.67	40.79
	Our model	**43.33**	**68.33**	**80.00**	**46.14**
EmoMV-C	Baseline (Modified model [5])	27.08	53.13	72.92	38.82
	Our model	**33.33**	**63.54**	**75.00**	**41.64**

Table 11: Given a music query, retrieve videos: Accuracy and the mAP on the EmoMV dataset collection.

Dataset	Multi-task learning models	Top-1(%)	Top-3(%)	Top-5(%)	mAP(%)
EmoMV-A	Baseline (Modified model [5])	25.20	49.60	76.80	39.45
	Our model	**56.00**	**83.60**	**90.80**	**52.31**
EmoMV-B	Baseline (Modified model [5])	43.17	70.83	75.83	43.72
	Our model	**44.17**	**77.50**	**86.67**	**46.83**
EmoMV-C	Baseline (Modified model [5])	39.58	76.04	85.42	38.59
	Our model	**46.88**	**75.00**	**86.46**	**41.98**

52.31%, respectively. These values obtained by using the baseline model are lower at 25.20%, and 39.45%, respectively. On the EmoMV-B and EmoMV-C datasets, using our model, the top-1 accuracy is 44.17% and 46.88%, respectively. The mAP score of the our model on these two datasets is 46.83% and 41.98%, respectively. These values on the EmoMV-B and EmoMV-C datasets are also higher than those obtained by using the baseline model, thus confirming that we have proposed a strong model for affective music-video retrieval.

6.2.3. Ablation Study

We perform an ablation study by removing the emotion classification branches from our affective music-video retrieval model as well as the baseline, and train them with the experimental setup described in Subsection 5.3.1. As shown in Tables 12 and 13, on all three datasets of the EmoMV collection, the single task learning models perform worse than the ones with multi-task learning. Additionally, whether doing multi-task or single task learning, our approach outperforms the baseline on the EmoMV collection. According to Table 12, for the task of retrieving music given videos as queries, our model performance decreases slightly after the emotion classification branches are removed, with the mAP declining from 58.53% to 50.41% while the top-1 accuracy does not change considerably on the EmoMV-A dataset, after removing the emotion classification branches. The mAP decreases from 46.14% to 43.50% (on the EmoMV-B dataset), and from 41.64% to 38.20% (on the EmoMV-C dataset). The top-1 accuracy and the mAP of the baseline model on the EmoMV-A dataset are significantly reduced from 34.40% to 20%, and 39.59% to 22.58%, respectively, after the music and video branches are removed from its structure. The performance of the baseline model on the EmoMV-B and EmoMV-C datasets also decreases, with the top-1 accuracy declining to 10.00%, and 7.29%, respectively. The mAP of the baseline model on the EmoMV-B dataset also declines from 40.79% to 25.73%. On the EmoMV-C dataset, this value decreases from 38.82% to 26.24%.

Table 12: Given a video query, retrieve music: Accuracy and the mAP on the EmoMV dataset collection with single task learning.

Dataset	Single task learning models	Top-1(%)	Top-3(%)	Top-5(%)	mAP (%)
EmoMV-A	Baseline (Modified model [5])	20	60	80	22.58
	Our model	**56.80**	**68.40**	**77.20**	**50.41**
EmoMV-B	Baseline (Modified model [5])	10.00	43.33	43.33	25.73
	Our model	**40.83**	**70.00**	**81.67**	**43.50**
EmoMV-C	Baseline (Modified model [5])	7.29	53.13	100	26.24
	Our model	**35.42**	**65.63**	**77.08**	**38.20**

As shown in Table 13, for the case of retrieving videos given music segments

Table 13: Given a music query, retrieve videos: Accuracy and the mAP on the EmoMV dataset collection with single task learning.

Dataset	Single task learning models	Top-1(%)	Top-3(%)	Top-5(%)	mAP (%)
EmoMV-A	Baseline (Modified model [5])	20	20	20	31.28
	Our model	**62.40**	**88.80**	**93.20**	**49.81**
EmoMV-B	Baseline (Modified model [5])	25.83	55.83	64.17	27.21
	Our model	**41.67**	**73.33**	**84.17**	**42.63**
EmoMV-C	Baseline (Modified model [5])	26.04	54.17	70.83	30.01
	Our model	**36.46**	**72.92**	**87.50**	**39.42**

as queries, after removing the emotion classification branches, the performance of our model and the baseline gets worse on all three datasets of the EmoMV collection. In particular, the top-1 accuracy and the mAP of our model decline to 62.40% and 49.81%, respectively, on the EmoMV-A dataset. These values are only 20% and 31.28% for the baseline model. Similarly, on the EmoMV-B and EmoMV-C datasets, the top-1 accuracy of our model is reduced to 41.67% and 36.46%, respectively, while the mAP declines to 42.63% and 39.42%, respectively. When the emotion classification branches are removed, the baseline model performs worse on the EmoMV-B and EmoMV-C datasets, with the top-1 accuracy of 25.83%, and 26.04%, respectively. On these datasets, the mAP of the baseline model is only 27.21% and 30.01%, respectively.

7. Conclusion

In this study, we tackle the problem of limited available data for affective audio-visual correspondence learning by constructing the EmoMV collection consisting of three datasets (EmoMV-A, EmoMV-B, and EmoMV-C). In these datasets, music and video streams are labelled as “matched” or “mismatched” in terms of the emotions they are conveying. This collection of datasets, together with the code of all our models is available online [5]. The EmoMv-A dataset is created by making use of the available MVED dataset (which is primarily used for the emotion classification task) with emotions annotated by humans. The EmoMV-B dataset is constructed by first manually selecting music video segments from the Music Mood dataset of the AudioSet ontology. The music and video modalities of these selected music video segments are then automatically labelled with emotion categories using a deep neural network (in particular, the modified Feature AttendAffectNet model). These music video segments together with the predicted emotion labels are then used to create the EmoMV-B dataset. In the EmoMV-C dataset, music video segments are first split from soundtrack music videos (of songs featured in movies) that we self-collected from YouTube, therefore, they might contain some movie scenes. This also makes this dataset different from others in the EmoMV collection. The modified Feature AttendAffectNet is also applied to automatically label the music and video streams of the soundtrack music video segments with emotion categories. These music video segments together with the predicted emotion labels are then used to create matched and mismatched music-video pairs. An online survey is then carried out to evaluate the accuracy of labels provided in our datasets. The survey results show that the matched and mismatched segments are mostly differentiated by humans. In addition, the overlapping rate between the predicted emotion labels provided in our datasets and the ones rated by humans is high. Notably, although the emotion classification task is not the main focus of this study, the modified Feature AttendAffectNet model outperforms other state-of-the-art approaches on this task on the MVED dataset.

In addition to the dataset creation, we also address the tasks of binary affective music-video correspondence classification and affective music-video retrieval. To tackle the former, a deep neural network structure is proposed to classify whether music-video pairs are matched or mismatched in terms of emotions. We use state-of-the-art pretrained deep neural networks as feature extractors

[5] The_link_will_be_added_after_the_peer_review_process

to obtain visual and audio features from video and music streams. Such visual and audio features are then projected into a common representation space, from which they are fused together, and a binary classification task (with “matched” or “mismatched” output) is carried out. The model is then trained in a multi-task learning manner, whereby the binary classification (for the ”matched” and ”mismatched” labels) together with emotion classification on the video and music streams are conducted simultaneously. To adapt our proposed model to the affective music-video retrieval task, we compute the cosine distance between the visual and audio embeddings, instead of fusing them together. Ablation studies are then carried out, whereby our multi-task learning model is converted to the one with single task learning by removing the emotion classification tasks on the music and video streams. As a result, our proposed model outperforms state-of-the-art approaches on the EmoMV dataset collection whether single task learning or multi-task learning is conducted.

In sum, in this study, we construct a collection of three publicly available ground-truth datasets, which could be used by researchers to explore the relatively unexplored tasks of affective audio-visual correspondence learning. In addition, we offer a strong benchmark model (with single-task and multi-task learning) together with results for each of our three created datasets. In future work, we might develop other deep neural networks and evaluate their performance on the EmoMV dataset collection. In addition, it would be good to conduct large-scale user studies to further verify the accuracy of labels offered in our newly created datasets and our proposed affective music-video retrieval model.

Acknowledgement

This work was supported by MOE2018-T2-2-161 and the SUTD President’s Graduate International Fellowship.

References

[1] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, R. He, Deep audio-visual learning: A survey, International Journal of Automation and Computing 18 (3) (2021) 351–376.

[2] R. Arandjelovic, A. Zisserman, Look, listen and learn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617.

[3] D. Surís, A. Duarte, A. Salvador, J. Torres, X. Giró-i Nieto, Cross-modal embeddings for video and audio retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

[4] S. Zhao, Y. Li, X. Yao, W. Nie, P. Xu, J. Yang, K. Keutzer, Emotion-based end-to-end matching between image and music in valence-arousal space, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2945–2954.

[5] B. Li, A. Kumar, Query by video: Cross-modal music retrieval., in: ISMIR, 2019, pp. 604–611.

[6] S. Horiguchi, N. Kanda, K. Nagamatsu, Face-voice matching using cross-modal embeddings, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1011–1019.

[7] A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: Cross-modal biometric matching, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8427–8436.

[8] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, W. Matusik, On learning associations of faces and voices, in: Asian Conference on Computer Vision, Springer, 2018, pp. 276–292.

[9] R. Arandjelovic, A. Zisserman, Objects that sound, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 435–451.

[10] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, A. Torralba, The sound of pixels, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 570–586.

[11] O. Slizovskaia, E. Gómez Gutiérrez, G. Haro Ortega, Correspondence between audio and visual deep models for musical instrument detection in video recordings (2017).

[12] R. Su, F. Tao, X. Liu, H. Wei, X. Mei, Z. Duan, L. Yuan, J. Liu, Y. Xie, Themes informed audio-visual correspondence learning, arXiv preprint arXiv:2009.06573 (2020).

[13] A. Nagrani, S. Albanie, A. Zisserman, Learnable pins: Cross-modal embeddings for person identity, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 71–88.

[14] E. Meyer, Meaning in music, Chicago: University of Chi-cago Press. Meyer1956Emotion and Meaning in Music (1956).

[15] Y. R. Pandeya, B. Bhattarai, J. Lee, Music video emotion classification using slow–fast audio–video network and unsupervised feature representation, Scientific Reports 11 (1) (2021) 1–14.

[16] P. Chua, D. Makris, D. Herremans, G. Roig, K. Agres, Predicting emotion from music videos: exploring the relative contribution of visual and auditory information to affective responses, arXiv preprint arXiv:2202.10453 (2022).

[17] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE transactions on affective computing 3 (1) (2011) 18–31.

[18] G. Verma, E. G. Dhekane, T. Guha, Learning affective correspondence between music and image, in: ICASSP 2019-2019 IEEE International Con-

ference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3975–3979.

[19] Y. R. Pandeya, B. Bhattarai, J. Lee, Deep-learning-based multimodal emotion classification for music videos, Sensors 21 (14) (2021) 4927.

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

[21] A. S. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients, Proceedings of the National Academy of Sciences 114 (38) (2017) E7900–E7909.

[22] P. N. Juslin, P. Laukka, Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening, Journal of new music research 33 (3) (2004) 217–238.

[23] L. Jaquet, B. Danuser, P. Gomez, Music and felt emotions: How systematic pitch level variations affect the experience of pleasantness and arousal, Psychology of Music 42 (1) (2014) 51–70.

[24] C. E. Osgood, W. H. May, M. S. Miron, M. S. Miron, Cross-cultural universals of affective meaning, Vol. 1, University of Illinois Press, 1975.

[25] P. J. Lang, Cognition in emotion: Concept and action, Emotions, cognition, and behavior 191 (1984) 228.

[26] P. Ekman, Basic emotions, Handbook of cognition and emotion 98 (45-60) (1999) 16.

[27] G. Colombetti, From affect programs to dynamical discrete emotions, Philosophical Psychology 22 (4) (2009) 407–425.

[28] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (6) (1980) 1161.

[29] P. J. Lang, M. M. Bradley, B. N. Cuthbert, Emotion, attention, and the startle reflex., Psychological review 97 (3) (1990) 377.

[30] Y. Baveye, E. Dellandrea, C. Chamaret, L. Chen, Liris-accede: A video database for affective content analysis, IEEE Transactions on Affective Computing 6 (1) (2015) 43–55.

[31] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, P. Maragos, Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization, EURASIP Journal on Image and Video Processing 2017 (1) (2017) 1–24.

[32] M. Soleymani, A. Aljanaki, Y. Yang, Deam: Mediaeval database for emotional analysis in music (2016).

[33] M. M. Bradley, M. K. Greenwald, M. C. Petry, P. J. Lang, Remembering pictures: pleasure and arousal in memory., Journal of experimental psychology: Learning, Memory, and Cognition 18 (2) (1992) 379.

[34] D. Watson, A. Tellegen, Toward a consensual structure of mood., Psychological bulletin 98 (2) (1985) 219.

[35] D. Watson, D. Wiese, J. Vaidya, A. Tellegen, The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence., Journal of personality and social psychology 76 (5) (1999) 820.

[36] P. Ekman, E. R. Sorenson, W. V. Friesen, Pan-cultural elements in facial displays of emotion, Science 164 (3875) (1969) 86–88.

[37] M. Zentner, D. Grandjean, K. R. Scherer, Emotions evoked by the sound of music: characterization, classification, and measurement., Emotion 8 (4) (2008) 494.

[38] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, D. Turnbull, Music emotion recognition: A state of the art review, in: Proc. ismir, Vol. 86, 2010, pp. 937–952.

[39] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, P. Lamere, The million song dataset (2011).

[40] A. Aljanaki, F. Wiering, R. C. Veltkamp, Studying emotion induced by music through a crowdsourcing game, Information Processing & Management 52 (1) (2016) 115–128.

[41] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE transactions on multimedia 7 (1) (2005) 143–154.

[42] W. A. Van Tilburg, T. Wildschut, C. Sedikides, Nostalgia's place among self-relevant emotions, Cognition and Emotion 32 (4) (2018) 742–759.

[43] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, Advances in neural information processing systems 29 (2016).

[44] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).

[45] J. S. Chung, A. Zisserman, Lip reading in the wild, in: Asian conference on computer vision, Springer, 2016, pp. 87–103.

[46] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, Lip reading sentences in the wild, in: 2017 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, 2017, pp. 3444–3453.

[47] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition (2015).

[48] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset, arXiv preprint arXiv:1706.08612 (2017).

[49] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, et al., Moments in time dataset: one million videos for event understanding, IEEE transactions on pattern analysis and machine intelligence 42 (2) (2019) 502–508.

[50] B. Xing, K. Zhang, L. Zhang, X. Wu, J. Dou, S. Sun, Image–music synesthesia-aware learning based on emotional similarity recognition, IEEE Access 7 (2019) 136378–136390.

[51] J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, H.-M. Wang, The acousticvisual emotion guassians model for automatic generation of music video, in: Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 1379–1380.

[52] D. Li, N. Dimitrova, M. Li, I. K. Sethi, Multimedia content processing through cross-modal association, in: Proceedings of the eleventh ACM international conference on Multimedia, 2003, pp. 604–611.

[53] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[54] K. Choi, G. Fazekas, M. Sandler, Automatic tagging using deep convolutional neural networks, arXiv preprint arXiv:1606.00298 (2016).

[55] H. T. P. Thao, B. Balamurali, G. Roig, D. Herremans, Attendaffectnet–emotion prediction of movie viewers using multimodal fusion with self-attention, Sensors 21 (24) (2021) 8356.

[56] H. T. P. Thao, D. Herremans, G. Roig, Multimodal deep models for predicting affective responses evoked by movies., in: ICCV Workshops, 2019, pp. 1618–1627.

[57] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 445–450.

[58] Y. Yi, H. Wang, Q. Li, Affective video content analysis with adaptive fusion recurrent network, IEEE Transactions on Multimedia 22 (9) (2019) 2454–2466.

[59] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems 27 (2014).

[60] L. Liu, L. Shao, X. Li, K. Lu, Learning spatio-temporal representations for action recognition: A genetic programming approach, IEEE transactions on cybernetics 46 (1) (2015) 158–170.

[61] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.

[62] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.

[63] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al., Ava: A video dataset of spatio-temporally localized atomic visual actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6047–6056.

[64] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, in: European Conference on Computer Vision, Springer, 2016, pp. 510–526.

[65] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, arXiv preprint arXiv:1808.01340 (2018).

[66] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[67] B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, Yaafe, an easy to use and efficient audio feature extraction software., in: ISMIR, Citeseer, 2010, pp. 441–446.

[68] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., Cnn architectures for large-scale audio classification, in: 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2017, pp. 131–135.

[69] Y. Yi, H. Wang, Multi-modal learning for affective content analysis in movies, Multimedia Tools and Applications 78 (10) (2019) 13331–13350.

[70] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, et al., Hear 2021: Holistic evaluation of audio representations, Proceedings of Machine Learning Research (PMLR) (2022).

[71] K. W. Cheuk, Y.-J. Luo, B. Balamurali, G. Roig, D. Herremans, Regression-based music emotion prediction using triplet neural networks, in: 2020 international joint conference on neural networks (ijcnn), IEEE, 2020, pp. 1–7.

[72] W. Jiang, Z. Wang, J. S. Jin, X. Han, C. Li, Speech emotion recognition with heterogeneous feature unification of deep neural network, Sensors 19 (12) (2019) 2730.

[73] D. Zeng, Y. Yu, K. Oyama, Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16 (3) (2020) 1–23.

[74] T. Rahman, M. Yang, L. Sigal, Tribert: Full-body human-centric audio-visual representation learning for visual sound separation, arXiv preprint arXiv:2110.13412 (2021).

[75] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[76] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, arXiv preprint arXiv:1606.01847 (2016).

[77] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Computer Vision and Image Understanding 174 (2018) 33–42.

[78] M. B. Jabra, R. Guetari, A. Chetouani, H. Tabia, N. Khlifa, Facial expression recognition using the bilinear pooling., in: VISIGRAPP (5: VISAPP), 2020, pp. 294–301.

[79] X. Wang, C. Yu, Y. Gu, M. Hu, F. Ren, Multi-task and attention collaborative network for facial emotion recognition, IEEJ Transactions on Electrical and Electronic Engineering 16 (4) (2021) 568–576.

[80] X. Jiang, L. Meng, D. Wu, Multi-task active learning for simultaneous emotion classification and regression, in: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2021, pp. 1947–1952.

[81] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[82] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[83] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: European conference on information retrieval, Springer, 2005, pp. 345–359.

[84] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, S. Drăghici, Machine learning and its applications to biology, PLoS computational biology 3 (6) (2007) e116.

[85] C. D. Brown, H. T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemometrics and Intelligent Laboratory Systems 80 (1) (2006) 24–38.

[86] D. Selva Ruiz, D. Fénix Pina, et al., Soundtrack music videos: The use of music videos as a tool for promoting films (2021).

[87] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016).

[88] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 1735–1742.

[89] E. Parisotto, J. L. Ba, R. Salakhutdinov, Actor-mimic: Deep multitask and transfer reinforcement learning, arXiv preprint arXiv:1511.06342 (2015).

[90] A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, R. Hadsell, Policy distillation, arXiv preprint arXiv:1511.06295 (2015).

[91] D. Zeng, Y. Yu, K. Oyama, Audio-visual embedding for cross-modal music video retrieval through supervised deep cca, in: 2018 IEEE International Symposium on Multimedia (ISM), IEEE, 2018, pp. 143–150.

[92] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, W. M. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.

[93] A. Sharma, A. Kumar, H. Daume, D. W. Jacobs, Generalized multiview analysis: A discriminative latent space, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 2160–2167.

[94] S. Teufel, An overview of evaluation methods in TREC ad hoc information retrieval and TREC question answering, in: Evaluation of Text and Speech Systems, 2007, Ch. 6, pp. 163–186. doi:10.1007/978-1-4020-5817-2_6.