Romantic and Family Movie Database: Towards Understanding Human Emotion and Relationship via Genre-Dependent Movies

Po-Chien Hsu
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
pochienhsu@gapp.nthu.edu.tw

Jeng-Lin Li
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
cllee@gapp.nthu.edu.tw

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Movie gains popularity by successfully immersing viewers into affective contents under an emotion perception and elicitation process. Film genre sets the tone of a movie to shape key emotional context during the storytelling procedure. In romantic and family movies, character relationship is a key contextual attribute guiding the whole storyline, which is expected to evoke a sense of being moved in the audience. In this work, we propose Romantic and Family Movie Database, which consists of 1029 movie clips from 10 romantic and family movies. We provide annotations including character relationship type, character relationship status, perceived emotions, induced emotions, and degree of being moved. Our analysis elaborates the inconsistency between perceived and induced valence with the support of annotations including character relationship status and degree of being moved. We also find that being moved and induced arousal are positively correlated in romantic and family movies. We present comprehensive baseline results in recognizing relationship status and emotions using different models and modalities. The database is publicly available at https://rfmd.ee.nthu.edu.tw/.**

***Index Terms*—movie database, film genre, character relationship, degree of being moved, perceived and induced emotions**

I. Introduction

With the advancement in video recording devices and large-volume data storage technology, multimedia contents of all kinds such as movies, TV series, short films, vlogs, and etc. have become ubiquitous. Meanwhile, an increasing number of viewers have spent more time experiencing videos via various platforms. Movie is renowned for reflecting human lives and societal phenomena through finely orchestrated audiovisual contents, and gains popularity due to its capability in triggering viewer emotions, making it a perfect medium as an affective stimuli. Moreover, movie-seeing itself acts as a spontaneous process of emotion perception and elicitation in the audience. Instead of adopting short videos uploaded by multimedia platform users or designated excerpts investigated by prior study [1], movies produced by professional filmmakers are more eligible for large-scale affective video content analysis. Creation of a movie database is a key and necessary step which provides data and ground truth labels to facilitate emotion recognition [2] and applications such as video indexing [3] and video summarization [4].

In the wide variety of movies, diverse film genres feature different types of story plots and character development, which are shaped by an underlying goal to evoke specific emotions in the audience [5], [6], e.g., horror movies are expected to evoke panic and fear. Such genre-specific movies effectively pile up the contextualized contents, where associated emotions frequently occur. A movie database can subsequently benefit from the realistic but relevant manifestation of contents and emotions. Considering film genres to consolidate target emotional contents avoids massive redundancy from movies. Previous movie databases providing emotional annotations without film genre selection [7], [8] make the affective content analysis inefficient. The messy annotated contents also lead to high complexity and difficulty in model training of video understanding [9], [10].

To systematically analyze human emotions in an into-life manner, interpersonal relationship serves as a contextualized attribute in movies, which is abundantly portrayed in romantic and family movies. Interpersonal relationship is bundled with in-movie context which projects the audience into social situations encountered by movie characters and thus induces audience emotions. In romantic and family movies, the audience falls deeper into the story with character development and corresponding interpersonal relationship that forms the feeling of being moved. For example, a strong sense of being moved will be evoked when the audience realizes the deep love between two young lovers who have cancer after watching them experience a series of events of joy and sorrow. Therefore, we identify character relationship as a key annotation to disclose human's affective appraisal mechanism of being moved and other emotional states driven by the interpersonal relationship.

In this work, we propose Romantic and Family Movie Database, which consists of 1029 movie clips excerpted from 10 romantic and family movies and provides annotations including character relationship type and status, perceived and induced emotions, and degree of being moved. Each clip was excerpted based on a complete interpersonal interaction. We

978-1-6654-5908-2/22/$31.00

TABLE I
MULTIMODAL MOVIE DATABASES WITH AUDIENCE EMOTIONAL ANNOTATION OR MOVIE CHARACTER ANNOTATION

Database	Genre	Audience Emotional Annotation	Movie Character Annotation
FilmStim	N/A	induced emotion	-
MAHNOB-HCI	N/A	induced emotion	-
AMIGOS	N/A	induced emotion	-
LIRIS-ACCEDE	All	induced emotion	-
COGNIMUSE	All	perceived emotion, induced emotion	-
MovieGraph	All	-	attributes, interactions, relationships, timestamps
MovieNet	All	-	bounding box, identity
Ours	Romance, Family	perceived emo., induced emo., being moved	relationship type, relationship status

explicitly define annotating rules for the *interaction-oriented movie clips* with the underlying character relationship status which is categorized into negative, neutral, or positive. Perceived and induced emotions of each clip were both labeled with valence and arousal dimensions. The degree of being moved was annotated together with induced emotions. In data analysis, the property of each annotation is explored by inter-annotator agreement, label distribution, and Pearson correlation coefficient. We analyze the relationship between perceived and induced valence with support of annotations particularly related to romantic and family movies, i.e., character relationship status and degree of being moved, which makes it possible to explain that the inconsistency of these two emotions often occurs in specific scenarios. Besides, we find that being moved and induced arousal are relevant in romantic and family movies. In baseline experiments, we report comprehensive results on different neural network models and modalities for character relationship status and audience emotions prediction. The database is publicly available at a website with downloadable files including multimodal feature sets, annotations, and start/end times of each movie clip.

II. RELATED WORK

A. Movie Databases

Creation of a movie database is a key and necessary step to acquire data and annotations which are able to effectively facilitate computational work. As shown in Table I, previous movie databases pick movies without genre selection, making it impossible to exploit contextual contents based on characteristics of specific film genres. FilmStim [11] selected movies from emotional categories including fear, anger, sadness, disgust, amusement, tenderness, and neutral. MAHNOB-HCI [12] categorized movie excerpts into disgust, amusement, joy, fear, sadness, and neutral. AMIGOS [13] selected movies from the top rated movie list in IMDb[1]. Other databases claim that their selected movies come from various film genres, e.g., LIRIS-ACCEDE [7] contains movies from up to 9 film genres. Most movie databases which provides emotional annotations only annotated induced emotions, such as FilmStim, MAHNOB-HCI, AMIGOS, and LIRIS-ACCEDE, while COGNIMUSE [8] annotated both perceived and induced emotions. However, none of them offers additional emotional annotations according to film genres. On the other hand, video understanding makes use of character information and aesthetic elements in movie databases. MovieGraph [9] annotated each clip with graph-based annotations of social situations in movie characters. MovieNet [10] provide annotations including character bounding box and identity, scene boundary, description sentence, place, action, and cinematic style. Despite the fact that both of them annotated a variety of movie character annotations, none of them highlights any higher-level annotation between characters. To the best of our knowledge, there has been no previous movie databases focusing on specific film genres or providing both audience emotional annotations and movie character annotations

[1]https://www.imdb.com/

B. Perceived Emotions versus Induced Emotions

When experiencing affective contents such as music or movie, the audience perceives the emotions conveyed by affective contents, which simultaneously induce emotional reaction in the audience. Perceived emotions refer to the emotions expressed by affective contents and recognized by the audience with objective criteria. In contrast, induced emotions represent the subjective emotions evoked in the audience, which is related to personal experience and individual preference [14]. Previous work suggests that perceived emotions are more objective compared to induced emotions [15], and annotators usually have stronger agreement on perceived emotions [16]. The relationship between these two emotions has been widely studied in music [16]–[18], while there only exists limited work on the correlation and difference between perceived and induced emotions in movies. Muszynski et al. [19] extended LIRIS-ACCEDE database [7] by annotating perceived emotions, and found that perceived and induced emotions are not always consistent. However, the conclusion is only supported by statistics such as Pearson correlation coefficient, without providing detailed analysis or logical explanation.

III. ROMANTIC AND FAMILY MOVIE DATABASE

Romantic and Family Movie Database consists of 1029 movie clips and provides audience emotional annotations and character relationship annotations. We specify the romantic and family movies as the most relevant genres with a collection of affective contents centered around interpersonal interaction. The proposed database contains 10 romantic and family movies with high IMDb ratings. Each movie clip

TABLE II
THE META INFORMATION OF THE 10 SELECTED ROMANTIC AND FAMILY MOVIES

Movie	Year	Genre	IMDb Rating	Character Relationship Type of Excerpts
About Time	2013	Comedy, Drama, Fantasy, Romance, Sci-Fi	7.8	Romance, Family
Me Before You	2016	Drama, Romance	7.4	Romance
The Theory of Everything	2014	Biography, Drama, Romance	7.7	Romance
The Fault in Our Stars	2014	Drama, Romance	7.7	Romance
The Spectacular Now	2013	Comedy, Drama, Romance	7.1	Romance
Lady Bird	2017	Comedy, Drama	7.4	Romance, Family, Friendship
Wonder	2017	Drama, Family	8.0	Family, Friendship
The Blind Side	2009	Biography, Drama, Sport	7.6	Family
Gifted	2017	Drama	7.6	Romance, Family
The Judge	2014	Crime, Drama	7.4	Romance, Family

was excerpted based on a complete interpersonal interaction between movie characters and annotated with character relationship status, i.e. whether characters get along well with each other. Perceived and induced emotions were annotated by two different sets of participants, respectively. In response to the essential purpose of romantic and family movies, i.e. inducing a sense of being moved in the audience, the degree of being moved was also annotated.

A. Data Collection Procedure

We include 10 popular romantic and family movies which have been in theater and received high IMDb ratings (not lower than 7.1) in the proposed database. Table II lists the selected movies with their released years, genre information provided by IMDb, and character relationship types of excerpted clips. *About Time, Me Before You, The Theory of Everything, The Fault in Our Stars, The Spectacular Now* are categorized as romantic movies according to their storylines and IMDb genres, while *Lady Bird, Wonder, The Blind Side, Gifted, The Judge* are categorized as family movies according to their storylines. Each movie is categorized into one genre based on the major storyline, while its excerpted clips might contain multiple character relationship types including romance, family, and friendship, because various story plots and character interactions are developed in a movie.

We set up rules to extract interaction-oriented excerpts belonging to three relationship types (romance, family, and friendship) from the 10 movies. Each movie clip was excerpted based on a complete interpersonal interaction, which is the basic semantic unit for comprehending and experiencing a story plot [20]–[22]. We define an interaction-oriented movie clip by the following criteria.

- A clip must contain an interpersonal interaction between two or more than two characters.
- The interpersonal interaction in a clip must be related to relationship types of romance, family, or friendship.
- A clip must be ended or segmented if a narrative shift occurs, i.e. the movie plot transits from one scene to another by shift in characters, shift in location, or shift in time [23].

There were 1029 movie clips excerpted according to the definition of an interaction-oriented movie clip, including 422 clips of romantic relationship, 499 clips of family relationship, and 108 clips of friendship relationship. The clips comprise various forms of interaction and result in varied length of contents ranging from 3 to 137 seconds. The average and median length of the clips are 26 seconds and 23 seconds, respectively. For each movie clip, its character relationship type and start/end times with respect to the timecode of original full movie are publicly available.

B. Multimodal Features

We extract features to describe the multimodal contents for computational modeling. We incline to capture character behavior and background contents in audio-visual features sets. Specifically, eGeMAPS [24] and VGGish [25] are acoustic feature sets using hand-crafted descriptors and a pretrained model. Meanwhile, Facial Action Units [26] and VGG19 [27] feature sets are low-level facial descriptors and general image recognition pre-trained embeddings. Movie subtitles provide a source of text modality indicating conversation transcripts and event cues and we extract word embeddings from a pre-trained BERT [28] model.

1) Audio: The extended Geneva minimalistic acoustic parameter set (eGeMAPS) extracted via openSMILE toolkit [24], [29] is developed for speech emotion recognition and behavior traits analysis. There are 23 acoustic low-level descriptors such as frequency, energy, and spectral related parameters which are encoded by statistical functionals as 88-dimensional feature vectors. VGGish is a pre-trained VGG-like audio classification model [25] used to capture background acoustic scenes and produces a 128-dimensional embedding. The pre-trained model is based YouTube-100M, a dataset consisted of 100 million YouTube videos. For each movie clip, time-series eGeMAPS feature set of every second and VGGish feature set of every 0.96 second are publicly available.

2) Visual: Facial Action Units (AUs) are generated to quantitatively describe facial expressions of movie characters. We apply OpenFace 2.0 toolkit to detect 17 AUs described by their intensities [26], [29], ranging from 0 to 5 with continuous values in between, which yields a 17-dimensional feature set. The 19-layer VGG network with batch normalization [27] pre-trained on ImageNet database (1.3M images) is applied to illustrate visual backgrounds in movies. Each frame in a movie clip is downsized to 224x224 as an input, and we extract

the 4096-dimensional embedding from the output of the last hidden layer. For each movie clip, which is in 30 fps, time-series AU feature set and pre-trained VGG19 feature set for every video frame, are publicly available.

3) Text: Movie subtitles not only contain transcripts of character conversation, but also include event cues such as *exhale, whoops, all laughing, girls giggling*, and etc. We utilize the pre-trained BERT-base-uncased model [28] to extract latent embeddings from text modality. Since each movie clip varies in word count of subtitles, the input sequence is truncated and padded to equal length, resulting in a 372-long and 768-dimensional embedding. For each movie clip, its English subtitles and pre-trained BERT features are publicly available.

C. Data Annotation

1) Character Relationship Status: Development of movie characters and interaction between them guide the plot and affective events in romantic and family movies. Therefore, we design to annotate character relationship status which concretizes an underlying state bundling humans together in an interaction-oriented movie clip. Common social norms shared by human beings allow annotators to objectively identify relationship status by observing character interaction and the corresponding relationship type. Annotators are asked to assess each clip independently for the character relationship status with 3 categories including negative, neutral, and positive, which is defined by the following criteria.

- If a viewer can tell that characters get along badly by their interactive behavior, the annotation should be labeled as negative. For example, friends have different opinions on an issue and start to curse each other.
- If a viewer can tell that characters get along well by their interactive behavior, the annotation should be labeled as positive. For example, lovers embrace each other warmly and talk in romantic words.
- If characters do not reveal their relationship status through interactive behavior in the clip or it cannot be classified to positive or negative, the annotation should be labeled as neutral. For example, a mother and her daughter listen to the poems recited on the radio together.

2) Perceived Emotions: Emotion perception refers to the abilities of recognizing and identifying emotions in others. When watching a movie, the audiences perceive and recognize the emotion conveyed by movie contents composed of character interaction, music, scenery, and etc. The annotation of perceived emotions contain both valence and arousal labels. Perceived valence was labeled with categorical scores scaled from 1 to 5, describing the degree of valence from negative to positive, while perceived arousal was labeled with categorical score in 3 levels, describing the degree of arousal from low to high. We recruited 8 annotators including 4 males and 4 females from an university to watch movies clips and annotate their perceived emotions. During the annotating process, they were asked to assess *What emotion does the movie clip convey?* or *How would you describe the affective content of the movie clip?*.

3) Induced Emotions: Induced emotions are meant to capture the audience emotions in response to the experienced contents after a series of cognitive and affective appraisal processes. After perceiving movie contents, the audience emotion will be induced by associating with personal experience and individual preferences. The annotation of induced emotions contain both valence and arousal labels. Induced valence was labeled with categorical scores scaling from 1 to 5, describing the degree of valence from negative to positive, while the induced arousal was also labeled with categorical scores scaling from 1 to 5, describing the degree of arousal from low to high. We recruited another set of 8 annotators (different from those who annotated perceived emotions) including 4 males and 4 females from an university to annotate induced emotions of each movie clip. First, annotators attended a screening session to watch the full movie, so that they understood the whole story and the developing direction of plots. On the second day, they could focus on labeling induced emotions of each movie clip from the movie they watched in the previous day. During the annotating process, they were asked to assess *what is your emotion when watching the movie clip?*.

4) Being Moved: Interaction-oriented movie clips collected from romantic and family movies are naturally endowed with a goal to make audiences being moved. Therefore, the degree of being moved indicates the intensity of induced emotional response from romantic and family movies. We quantize the emotional degree of being moved in a 5-level categorical scale under the annotating procedure of induce emotions by the same group of 8 annotators described in Section III-C3.

For each movie clip, we publicly release its annotations including character relationship status, perceived emotions, induced emotions, and degree of being moved.

IV. Data Analysis

We calculate inter-annotator agreement of each annotation, showing that being moved is a more concordant and concrete annotation than induced arousal in romantic and family movies. The label distribution presents different influences of emotion perception and elicitation. Pearson correlation coefficients provide an insight on the relevance in two sets of annotations, one including relationship status, perceived and induced valence, while the other including induced arousal and being moved. Moreover, we demonstrate how relationship status and being moved as indicators help bridge the gap between perceived and induce emotions.

A. Inter-Annotator Agreement

Inter-annotator agreement is a metric used to evaluate the degree of agreement over multiple annotators when annotating labels with subjectivity. Different levels of subjectivity exist in emotional annotations and thus we adopt four measures including percent agreement, Fless' kappa [30], Randolph's kappa [31], and Krippendorff's alpha [32] to evaluate annotations in the proposed database. Percent agreement is the most intuitive but the least robust measure, which does not take into account the fact that agreement may happen by chance [33].

TABLE III
INTER-ANNOTATOR AGREEMENT

Measure	Per-V	Per-A	Ind-V	Ind-A	Moved
Percent Agreement	0.437	0.540	0.434	0.299	0.372
Fless' Kappa	0.285	0.216	0.208	0.016	0.042
Randolph's Kappa	0.296	0.310	0.293	0.124	0.215
Krippendorff's Alpha	0.670	0.335	0.487	0.093	0.186

Note: Per-V stands for *perceived valence*; Per-A stands for *perceived arousal*; Ind-V stands for *induced valence*; Ind-A stands for *induced arousal*; Moved stands for *being moved*.

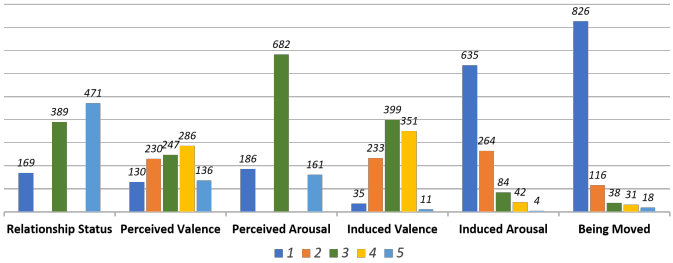

Fig. 1. *Label distribution of each annotation.*

Fleiss' kappa is a widely used measure capable of removing the chance agreement, a numerical term if all annotators give their labels in random. Randolph's free-marginal multirater kappa fixes the limitation that Fleiss' kappa is influenced by prevalence and bias. It is appropriate to apply this measure to annotations with unbalanced label distribution. Krippendorff's alpha is estimated depending on disagreement instead of agreement used in kappa statistics, and also considers disagreement expected by chance. Percent agreement ranges from 0 to 1, while kappa and alpha statistics range from -1 to 1, where negative values indicate systematic disagreement.

Table III shows the inter-annotator agreement of 5 annotations with different measures, where perceived arousal is annotated with 3 levels while other annotations are in 5 levels. We can observe that perceived emotions have higher agreement than induced emotions, which corresponds to the fact that perceived emotions are more objective. Besides, valence emotions have higher agreement than arousal emotions, where previous work also shows similar results [8]. Among low agreement annotations such as induced arousal and being moved, all measures in Table III indicate that being moved is more concordant than induced arousal, both of which are annotated with 5 levels by the same group of 8 annotators. The results imply that it is easier for the audience to recognize being moved than induced arousal because romantic and family movies are designed with an aim to evoke a sense of being moved. That is, the emotional intensity in events and stories particularly built upon romantic and family relationships are more adequately to be quantized using a sense of being moved. In this case, induced arousal is general but less concrete in the assessment of particular scenarios.

B. Label Distribution

The label distribution of each annotation is shown in Fig. 1. Annotations labeled in 3 levels such as character relationship status and perceived arousal are presented in categorical scores of (1, 3, 5), while those labeled in 5 levels are presented in (1, 2, 3, 4, 5). Character relationship status was labeled on a basis of strict definitions executed in several rounds of delicate discussions until achieving an agreement between annotators. Other labels on each clip are aggregated using majority voting over the raw scores from 8 annotators. In Fig. 1, we do not consider samples with more than one labels receiving the maximum number of votes, where 152 samples in perceived valence, 126 samples in perceived arousal, 135 samples in induced valence, 200 samples in induced arousal, and 148 samples in being moved are excluded, respectively.

In Fig. 1, we can observe that character relationship status has more positive labels than negative or neutral. The proposed database focusing on romantic and family movies presents abundant pleasant and warm movie plots and gives rise to numerous clips with positive interpersonal relationship. Besides, neutral labels also account for a large proportion, representing diverse contents of plain scenes in daily lives. Perceived valence is more evenly distributed than induced valence. The phenomenon that much fewer samples categorized into extreme classes (label 1 or 5) indicates potential inconsistency between perceived and induced emotions. Induced emotions after mental assessment of the audience would not be as strong as perceived emotions directly created by the content stimuli. On the other hand, perceived arousal has more labels in large values than induced arousal, which leads to the similar emotion perception and elicitation conclusion from the analysis of valence. Both induced arousal and being moved have a declining number of samples as label value increases.

C. Pearson Correlation Coefficient

As shown in Table IV, we calculate Pearson correlation coefficient (PCC), which is a measure of linear correlation between two sets of data, to find the pairwise relationship among all the annotations. We use the same aggregated labels resulting from the majority vote of 8 annotators but assign the average value for the situation that a sample obtain more than one highest voted labels.

In Table IV, we can observe that perceived valence and induced valence have positive and high PCC of 0.720, while perceived arousal and induced arousal have positive but low PCC of 0.243. It suggests that in the comparison between perceived and induced emotions, the consistency in arousal is weaker than that in valence. PCC between character relationship status and perceived valence achieves the highest value of 0.739, indicating that these two annotations are conceptualized in a similar line of movie content presentation. Besides, character relationship status and induced valence also share a positive PCC of 0.576, reflecting the connection between movie characters and the audience. Last but not least, PCC between induced arousal and being moved receives a positive value of 0.413, suggesting that the evoked emotional states have shared levels of intensity in the context of romantic and

TABLE IV
PAIRWISE PEARSON CORRELATION COEFFICIENT

Annotation	Per-V	Per-A	Ind-V	Ind-A	Moved
Relationship	**0.739**	-0.311	**0.576**	-0.076	0.177
Per-V	∗	-0.221	**0.720**	-0.101	0.096
Per-A	∗	∗	-0.095	0.243	-0.093
Ind-V	∗	∗	∗	-0.075	0.007
Ind-A	∗	∗	∗	∗	**0.413**

Note: Relationship stands for *character relationship status*; Per-V stands for *perceived valence*; Per-A stands for *perceived arousal*; Ind-V stands for *induced valence*; Ind-A stands for *induced arousal*; Moved stands for *being moved*.

family stories. For example, warm or heart-breaking movie plots involving the audience into a strong feeling of high arousal situation would be interpreted as being moved, when experiencing romance and family oriented scenes. The findings suggest that the annotation of being moved is a promising indicator to describe a similar but more consistent emotional aspect than the annotation of induced arousal in the romance and family related scenarios.

D. Perceived Valence versus Induced Valence

In Table IV, PCCs between perceived emotion and induced emotion show that these two emotions are more correlated in valence (PCC=0.720) than in arousal (PCC=0.243), and previous work yields similar results [19]. It indicates that perceived and induced valence values are consistent in most movie clips, while opposite valence values in these emotions can be considered a special case. We try to provide a detailed analysis in such condition with the support of newly collected annotations particularly for romantic and family movies, i.e. character relationship status and being moved.

Since both perceived valence and induced valence are labeled in 5 levels, a score equal to or larger than 4 is regarded as a positive emotion, while a score equal to or smaller than 2 is regarded as a negative emotion. Among 1029 movies clips, there are 12 clips labeled as positive perceived valence and negative induce valence. The character relationship status of these clips presents 0 negative labels, 2 neutral labels, and 10 positive labels, suggesting that characters get along well in such scenario. On the other hand, the average being moved score of these clips is 2.71, which is larger than that of all clips, namely 1.47, indicating that the audience feels more moved when watching these clips. In summary, these clips illustrate a scenario where movie characters feel warm and cherish each other, displaying positive perceived emotion and positive character relationship status. However, the audience knows that characters have suffered from or going to suffered from hardship and demonstrate humanity virtue, thus inducing negative induced valence and high degree of being moved. Such scenario usually depicts a heart-breaking story plot about characters leaving apart or facing death.

In contrast, there are 22 clips labeled as negative perceived valence and positive induce valence. The character relationship status of these clips presents 7 negative labels, 11 neutral labels, and 4 positive labels. It suggests that characters get along badly in the scenario, or such clips simply present casual interaction in daily lives, which is unable to be recognized as either positive or negative relationship status. On the other hand, the average being moved score of these clips is 1.20, which is smaller than that of all clips, indicating that the audience feels less moved when watching these clips. In summary, these clips illustrate a scenario where characters are unkind and argue with each other, and some of them is simply an harmless quarreling, displaying negative perceived emotion and negative/neutral character relationship status. However, the audience knows that characters are not meant to harm each other and regards it as an funny and ridiculous plot, thus inducing positive induced valence and low degree of being moved. Such scenario usually depicts a comedy or a prank in characters' daily lives.

V. BASELINE EXPERIMENT

A. Experimental Setup

In this section, we establish baselines of binary classification tasks using different annotations, feature sets, and models on Romantic and Family Movie Database. The annotations are binarized into negative and positive classes as the prediction target. We regard samples with label values larger and equal to 3 as the positive class and other samples as the negative class. A previous study has shown the difference between neutral or positive emotion attributes is less significant [34]. The situation is obvious in romantic and family movies with many peaceful and harmonious scenes and thus we categorize them into the non-negative class. For the numbers of (negative, positive) labels in each annotation, there are (169, 860) labels in character relationship, (293, 689) labels in perceived valence, (87, 865) labels in perceived arousal, (190, 772) labels in induced valence, (800, 120) labels in induced arousal, and (847, 95) labels in being moved. For input features, we adopt features sets presented in Section III-B, including eGeMAPS, VGGish, Facial Action Units, pre-trained VGG19 with batch normalization, and pre-trained BERT. We investigate the recognition results from single-modal and multimodal models using audio, visual, and text. The multimodal model is implemented by late fusion concatenating features including VGGish, Facial Action Units, and pre-trained BERT. Experiments are conducted on the following neural network (NN) models:

- Dense: A deep network consists of 3 fully connected layers with all hidden nodes being 16-dimensional.
- LSTM: A network using a fully connected layer as input layer followed by an LSTM layer and two fully connected layers, where all hidden nodes are 16-dimensional.
- Self-Attention: A network consists of a fully connected layer followed by a Self-Attention layer and two fully connected layers, where all hidden nodes are 16-dimensional. Two heads are learned in the Self-Attention layer with a dropout rate 0.1.

Each layer of all the models uses ReLU as an activation function and the last layer uses Softmax for classification.

TABLE V
THE PERFORMANCE IN RECOGNIZING 6 TYPES OF ANNOTATIONS BY 6 TYPES OF FEATURE SETS AND 3 TYPES OF MODELS, RESPECTIVELY, PRESENTED IN MACRO F1-SCORE (MA. F1) AND WEIGHTED F1-SCORE (WT. F1).

Model	Modality	Feature Set	Relationship		Per-Valence		Per-Arousal		Ind-Valence		Ind-Arousal		Being Moved	
			ma. F1	wt. F1	ma. F1	wt. F1	ma. F1	wt. F1	ma. F1	wt. F1	ma. F1	wt. F1	ma. F1	wt. F1
Dense	Audio	eGeMAPS	0.566	0.755	0.549	0.621	0.611	0.846	0.484	0.636	0.574	0.759	0.557	0.789
	Audio	VGGish	0.564	0.747	0.559	0.628	0.627	0.856	0.573	0.693	0.554	0.737	0.584	0.795
	Visual	Action Units	0.552	0.763	0.607	0.666	0.475	0.863	0.560	0.685	0.561	0.758	0.528	0.787
	Visual	VGG19	0.566	0.751	0.553	0.624	0.579	0.846	0.561	0.718	0.498	0.731	0.534	0.815
	Text	BERT	0.670	0.809	0.726	0.768	0.605	0.850	0.678	0.786	0.604	0.807	0.649	0.852
	Multimodal	VGGish+AU+BERT	0.709	0.839	0.722	0.767	0.622	0.858	0.679	0.788	0.629	0.819	0.667	0.862
LSTM	Audio	eGeMAPS	0.579	0.738	0.525	0.572	0.606	0.820	0.562	0.711	0.533	0.745	0.591	0.815
	Audio	VGGish	0.617	0.770	0.578	0.630	0.652	0.853	0.610	0.728	0.486	0.722	0.619	0.826
	Visual	Action Units	0.442	0.728	0.614	0.651	0.476	0.865	0.511	0.743	0.465	0.809	0.473	0.851
	Visual	VGG19	0.517	0.730	0.542	0.619	0.512	0.815	0.551	0.700	0.451	0.725	0.524	0.808
	Text	BERT	0.646	0.795	0.691	0.740	0.591	0.841	0.678	0.789	0.602	0.804	0.622	0.848
	Multimodal	VGGish+AU+BERT	0.653	0.803	0.686	0.732	0.646	0.878	0.662	0.782	0.597	0.806	0.628	0.851
Self-Attention	Audio	eGeMAPS	0.612	0.755	0.561	0.615	0.614	0.815	0.542	0.680	0.525	0.736	0.570	0.789
	Audio	VGGish	0.605	0.768	0.556	0.614	0.664	0.861	0.591	0.718	0.548	0.740	0.595	0.815
	Visual	Action Units	0.565	0.714	0.628	0.673	0.483	0.851	0.561	0.684	0.485	0.780	0.514	0.820
	Visual	VGG19	0.540	0.736	0.550	0.625	0.522	0.822	0.536	0.697	0.478	0.733	0.528	0.801
	Text	BERT	0.662	0.807	0.707	0.753	0.583	0.845	0.660	0.773	0.608	0.813	0.627	0.855
	Multimodal	VGGish+AU+BERT	0.692	0.831	0.711	0.756	0.638	0.874	0.672	0.788	0.596	0.813	0.649	0.866

Each model is trained in 30 epochs with batch size of 16, Adam optimizer with learning rate of 0.0001, and cross entropy loss with class weights. The weight of majority class is 1.0, while that of minority class is the sample number of majority class divided by the sample number of minority class. We perform Leave-One-Movie-Out protocol to evaluate the proposed database, to prevent data contamination in model learning due to a specific style in audiovisual presentation generated from the same movie.

B. Experimental Results

In Table V, we can observe that the performance on both macro F1-score and weighted F1-score in predicting the degree of being moved is comprehensively better than that of induced arousal. For example, using multimodal features with Self-Attention model achieves 0.649 in macro F1-score and 0.866 in weighted F1-score in the task of predicting being moved, while that of induced arousal only obtains 0.596 and 0.813. That is, the degree of being moved is more predictable than the induced arousal for the recognition of romance and family related contents.

Among 5 single-modality feature sets, pre-trained BERT outperforms others, achieving the highest macro F1-score of 0.726 in predicting perceived valence using Dense model and the highest weighted F1-score of 0.855 in predicting being moved using Self-Attention model. The text modality, composed of movie subtitles, includes transcripts of character conversation and event cues, and thus acts as the most structured representation relevant to delivered relationship status and emotions. In contrast, the audiovisual representations can include redundancy in the complex presentations intertwined with various foreground and background elements such as human voice, music, character's action, scene and etc.

As shown in Table V, macro and weighted F1 scores of multimodal learning outperforms those of single-modal learning in most tasks using different models and annotations. The results demonstrate the effectiveness of combining multimodal representation in recognition tasks and also set up a benchmark for further development of research.

VI. CONCLUSION

In this work, we propose Romantic and Family Movie Database, which includes novel annotations such as character relationship status and degree of being moved on a collection of romance and family related movie clips. The identified annotations enable in-depth understanding of emotion perception and elicitation by the intermediate annotations of character relationship. We publicly release the multimodal movie database with 1029 movie clips from 10 movies for further research development. We also demonstrate the protocol that identifying a key annotation to characterize a genre of data can enhance affective video content analysis. In our data analysis results, the difference between perceived and induced emotions can be elaborated with the support of other annotations, and being moved has been shown to be a more concordant and predictable annotation than induced arousal in romantic and family movies.

For future work, we will further scale up the movie database using the proposed protocol and extend interpersonal relationship type to cover comprehensive social conditions such as friendship, relationships at work, and at school. The annotation of character relationship and being moved can be labeled for various multimedia data with human interaction contents. In addition, the protocol can be applied to the database construction of other film genres to shed a light on emotions using contextualized attributes.

REFERENCES

[1] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[2] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 396–409, 2018.

[3] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[4] S. hua Zhong, J. Wu, and J. Jiang, "Video summarization via spatio-temporal deep architecture," *Neurocomputing*, vol. 332, pp. 224–235, 2019.

[5] T. Grodal, "Moving pictures: A new theory of film genres, feelings, and cognition," 1999.

[6] G. M. Smith, *Film structure and the emotion system*. Cambridge University Press, 2003.

[7] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[8] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and 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*, vol. 2017, no. 1, pp. 1–24, 2017.

[9] P. Vicol, M. Tapaswi, L. Castrejón, and S. Fidler, "Moviegraphs: Towards understanding human-centric situations from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: A holistic dataset for movie understanding," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 709–727.

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

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[13] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[14] C. Plantinga, "Art moods and human moods in narrative cinema," *New Literary History*, vol. 43, no. 3, pp. 455–475, 2012.

[15] G. Matthews, D. M. Jones, and A. G. Chamberlain, "Refining the measurement of mood: The uwist mood adjective checklist," *British Journal of Psychology*, vol. 81, no. 1, pp. 17–42, 1990.

[16] Y. Song, S. Dixon, M. Pearce, and A. Halpern, "Perceived and induced emotion responses to popular music: Categorical and dimensional models," *Music Perception: An Interdisciplinary Journal*, vol. 33, pp. 472–492, 04 2016.

[17] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1_suppl, pp. 123–147, 2001.

[18] K. Kallinen and N. Ravaja, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, vol. 10, no. 2, pp. 191–213, 2006.

[19] M. Muszynski, L. Tian, C. Lai, J. D. Moore, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Recognizing induced emotions of movie audiences from multimodal information," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 36–52, 2021.

[20] L.-H. Chen, Y.-C. Lai, and H.-Y. M. Liao, "Movie scene segmentation using background information," *Pattern Recognition*, vol. 41, no. 3, pp. 1056–1065, 2008.

[21] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, "A local-to-global approach to multi-modal movie scene segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 143–10 152.

[22] I. U. Haq, K. Muhammad, T. Hussain, S. Kwon, M. Sodanil, S. W. Baik, and M. Y. Lee, "Movie scene segmentation using object detection and set theory," *International Journal of Distributed Sensor Networks*, vol. 15, no. 6, 2019.

[23] J. Cutting, "Narrative theory and the dynamics of popular movies," *Psychonomic Bulletin and Review*, vol. 23, pp. 1713–1714, 2016.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[25] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[26] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, 2015, pp. 1–6.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[29] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[30] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[31] J. J. Randolph, "Free-marginal multirater kappa (multirater k[free]): An alternative to fleiss' fixed-marginal multirater kappa." 2005.

[32] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] L. Fernández-Aguilar, B. Navarro-Bravo, J. Ricarte, L. Ros, and J. M. Latorre, "How effective are films in inducing positive and negative emotional states? a meta-analysis," *PLOS ONE*, vol. 14, pp. 1–28, 11 2019.