VACD: A Video Affective Dataset with Content Descriptions(Appendix)

Anonymous Author(s)

A Related Work

In this section, we will present and analyze other datasets with content similar to ours. In addition to the traditional video affective analysis datasets, which include videos and their corresponding emotion labels, we also observed some new multimodal datasets that include textual modalities and can be used for video affective analysis tasks. We will introduce both types of datasets in the following.

A.1 Datasets for Video Affective Content Analysis

Compared to datasets that require collecting physiological signals from video viewers, constructing a video affective analysis dataset does not require special equipment. After fully watching and carefully understanding the video content, viewers label the corresponding emotions, greatly simplifying the dataset creation process. Moreover, building a video affective analysis dataset does not necessitate that annotators work entirely in a laboratory environment. With proper supervision and inspection, researchers can more easily obtain a larger amount of valid data.

As a result, many datasets for annotating the affective content of movies have emerged. For example, the PMSZU dataset [**?**] consists of 386 video clips extracted from 8 popular movies, specifically designed for video affective content analysis. FilmStim [**?**] consists of various movie scenes where emotional labels are directly assessed by participants according to film mood and stylistic attributes. MediaEval2016 [**?**] is an extension of the LIRIS-ACCEDE dataset, adding 1,200 video clips as a test set to enhance research on affective content analysis. 50-Film [**?**] consists of various movie scenes where emotional labels are directly assessed by participants according to film mood and stylistic attributes. LIRIS-ACCEDE [**?**] consists of 9,800 video excerpts from 160 short and feature films, annotated for valence and arousal to facilitate affective content analysis. Compared to LIRIS-ACCEDE, the labels in our dataset, VACD, are *intended* emotion labels that annotators provided based on their perception and judgment after fully watching the movies. As emphasized in [**?**], emotion content based on viewers' perceptions is more suitable for computational analysis.

Additionally, user-generated videos are increasingly being used as raw data for video affective datasets. For instance, VideoEmotion [**?**] comprises annotated video clips where emotional labels are directly assessed by participants based on visual and auditory content. YF-E6 [**?**] consists of manipulated facial videos where labels are directly assessed by participants based on the realism and detectability of the forgery.

In Table 1, we can see that compared to these datasets, our dataset, VACD, introduces textual content related to the video, enriching the information contained within the dataset. Readers can refer to Section 3 of this paper for detailed information.

A.2 Video Caption for Affective Analysis

Since our dataset includes textual content related to the video, in addition to the aforementioned video affective analysis datasets, we must also introduce some datasets used for video caption tasks. For example, MSR-VTT [**?**] contains 10,000 YouTube video clips (10-30 seconds each) across 20 categories with 20 human-generated textual descriptions per video. It is used for video captioning, retrieval, and multimodal affective analysis, serving as a benchmark for video content research. The dataset includes mp4 format video files and json format annotation files, available via the Microsoft Research website.

The textual modalities in these datasets don't include emotional information. If multimodal fusion techniques are directly applied to these datasets for video affective analysis, the results may not be very effective. Consequently, datasets like EmVidCap[**?**], which introduce emotional content into video caption, have begun to emerge. However, this dataset primarily add adjectives or adverbs that describe emotions into the generated text, which only reflects the emotions of the characters in the video and does not always align with the emotions that the video creators intended to convey. The EMVPC dataset[**?**] expands upon EmVidCap by increasing the number of emotional words used and incorporating logical words tailored to the characteristics of the video.The primary application scenarios for these two datasets are video caption tasks, so they do not include clear affective labels corresponding to the video clips. In other datasets that include textual content and are used for video affective analysis tasks, such as the VAD dataset [**?**], the textual content often struggles to fully align with the video content.

VACD perfectly addresses the aforementioned issues. The textual modality in our dataset not only includes descriptions of the video content but also provides depictions of the characters' expressions and psychological states. These texts, created and carefully edited by professionals to aid visually impaired individuals in understanding the films, more effectively reflect the content and emotions that the films intend to convey. For a detailed comparison, readers can refer to Table 2.

B Detailed Feature Extraction Procedure

For the visual modality, each video is uniformly sampled at 8 frame increments to reduce the computational cost. Then, we extract 768-dimensional visual features using CLIP-large[**?**], which can capture universal representations. In the case of the audio modality, FFMpeg is employed to extract audio from the original movie video, with the audio format subsequently unified to 16 kHz. Subsequently, VGGish[**?**] is used to extract 128-dimensional audio features. With regards to the text modality, Tencent Cloud Automatic Speech Recognition (ASR) is employed to extract subtitles. Subsequently, the Chinese-MacBERT-large [**?**] model is employed to extract 1024-dimensional text features.

Table 1: Overview of Video Emotion Datasets

Dataset	Source	Scale	Annotations	Modality
PMSZU	Movie	386 clips	Arousal, genre of the movies	Audio, Visual
FilmStim	Movie	394 clips	Valence, arousal, 7 emotions	Audio, Visual
MediaEval2016	Movie	1200 clips	Valence, arousal	Audio, Visual
50-Film	Movie	2069 clips	Hedonic tone, energetic arousal, tense arousal	Text, Audio, Visual
LIRIS-ACCEDE	Movie	9,800 clips	Valence, arousal	Audio, Visual
VideoEmotion	User-generated	1101 clips	8 emotions	Audio, Visual
YF-E6	User-generated	1637 clips	6 emotions	Audio, Visual
VACD(ours)	Movie	19894 cilps	Valence, arousal, 10 emotions	Text, Audio, Visual

Table 2: Overview of Video Emotion Datasets with Text Annotations

Dataset	Source	Scale	Task	Annotations	Alignment
MSR-VTT	User-generated	10K clips	Video Caption	Action/Event descriptions	Yes
EmVidCap	User-generated	2K clips	Video Caption	Video descriptions containing emotional words	Yes
EMVPC	User-generated	1K clips	Video Caption	Video descriptions containing emotional and logical words	Yes
VAD	User-generated	19267 clips	Video Affective Analysis	Valence, arousal, valence comparison, arousal comparison, 13 emotions	No
VACD(ours)	Movie	19894 clips	Video Affective Analysis	Valence, arousal, 10 emotions	Yes

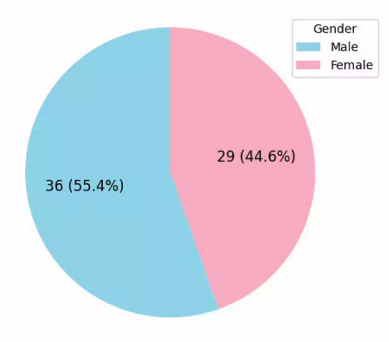

Figure 1: Gender distribution of annotators.

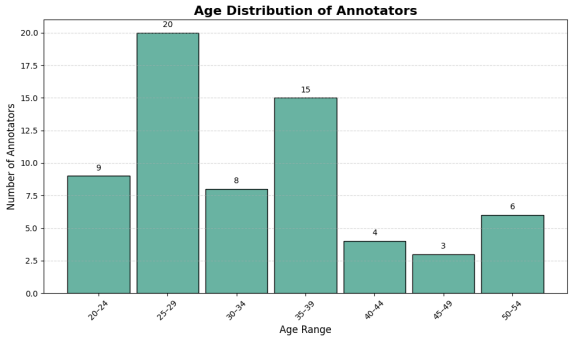

Figure 2: Age distribution of annotators.

C Baseline Models and Experiment setup

C.1 Baseline Models

AMP employs robust representation through the use of adversarial temporal masking and adversarial parameter perturbation. The former entails the introduction of an adversarial temporal masking strategy, which serves to enhance the encoding of other modalities. The latter, meanwhile, entails the introduction of an adversarial parameter perturbation strategy, which serves to enhance the model's generalization capabilities by adding an adversarial perturbation to the parameters of the model.

CARAT devises a reconstruction-based fusion mechanism to better model fine-grained modality-to-label dependencies by contrastively learning modal-separated and label-specific features. Finally, based on the reconstructed embeddings, they proposed a novel samplewise and modality-wise shuffle strategy to enrich the cooccurrence dependencies among labels.

TFN introduces a new model, termed Tensor Fusion Network, which learns both the intra-modality and inter-modality dynamics end-to-end. Inter-modality dynamics are modeled with a new

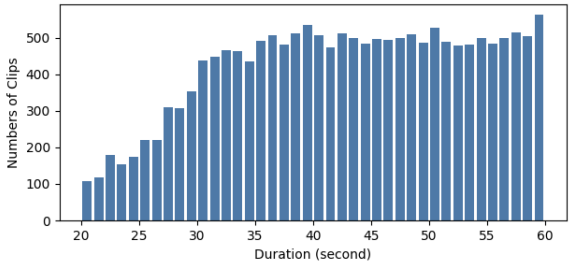

Figure 3: The duration distribution of clips.

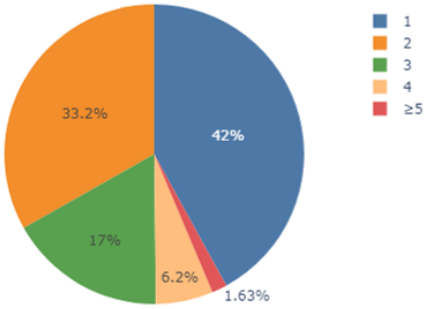

Figure 4: The proportion of video clips annotated with different numbers of emotion labels.

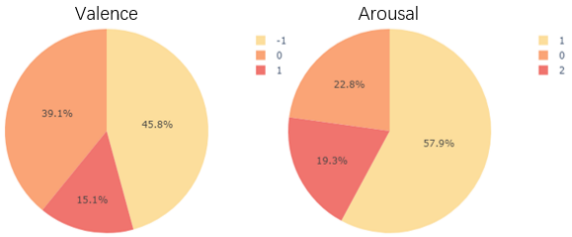

Figure 5: The statistical results of Valence and Arousal annotations in the dataset.

multimodal fusion approach, named Tensor Fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. Intramodality dynamics are modeled through three Modality Embedding Subnetworks, for language, visual and acoustic modalities, respectively.

C.2 Experiment setup

In this section, we formalize the analysis of affective content and provide detailed descriptions of the implementation methodologies for the three baseline approaches utilized in our experiments.

Let $X^m \in F^{t_m \times d_m}$ represents the feature of a modality. m can represent audio (a), visual (v) and text (t) in the dataset. t_m represents the temporal length of a modality, and d_m represents the feature dimension of a modality. Let $Y = \{y_1, y_2, \dots y_p\}$ represent the label space with p labels. For experiments about arousal and valence, p is equal to 3, while p is equal to 2 for the experiments about primary emotion. The task of classification is to predict the corresponding label y_i given the input features of i-th clip X_i^m, where $m \in \{a, v, t\}$.

The training, validation, and test sets are split with a ratio of 0.7, 0.1, and 0.2. We partition randomly our dataset five times and perform five replicate trials to enhance confidence for experimental results. All clips from the same movie in each division will only appear in one of the training, validation and test sets at the same time. Each split of the training set, validation set, and test set contains all label categories, with the label distribution proportions being roughly the same.

The specific parameter settings for each baseline are as follows. For AMP, in the PE experiment, the learning rate is set to 1e-5, weight decay to 1e-2, training for 30 epochs, and using AsymmetricLoss [**?**] as the classification loss function. In the VA experiment, the learning rate is set to 5e-5, weight decay to 1e-2, training for 20 epochs, and using weighted BCE Loss. In both cases, the batch size is 16, and the hidden feature dimension is 256. For CARAT, in the PE experiment, using AsymmetricLoss as the classification loss function. In the VA experiment, using weighted BCE Loss. In both cases, the batch size is 64, hidden feature dimension is 256, learning rate is set to 5e-5, weight decay to 1e-2 and training for 20 epoch.

For TFN, the learning rate is set to 1e-3, training for 30 epochs, weight decay to 1e-5. We use MSEloss as the classification loss function. The batch size is 32. All models were trained on NVIDIA GeForce RTX 3090.

For every emotion,we will get a regression value.And we set thresholds to classify.In the PE experiment,the threshold is set to [0.25,0.3,0.2,0.2,0.3,0.08,0.12,0.1,0.07,0.07].In the VA experiment, the threshold for V is set [-0.6,0.1] and the threshold for A is set [0.8,1.3].

D Statistical Charts of the Dataset

The gender and age distributions of the dataset annotators are shown in Figure 1 and Figure 2, respectively.

The histogram of video clip durations in the dataset is shown in Figure 3. The proportion of video clips annotated with different numbers of emotion labels is shown in Figure 4. The statistical results of Valence and Arousal annotations in the dataset are shown in Figure 5.

E Privacy and Copyright Considerations in Dataset Annotation

Although our dataset is constructed using publicly available movies from the web, the textual modality content was specifically developed for the platform, and no copyrighted movie content itself is used. Consequently, we will only publicly release the extracted visual, audio, and text features. We will periodically apply state-of-the-art feature extraction methods to reprocess the raw data and update the dataset accordingly.

During dataset construction, we collected only three demographic pieces of information from annotators: **age**, **gender**, and **educational background**. The annotators have explicitly **consented** to the use of this information for dataset construction and subsequent scientific research. This dataset is intended **solely** for scientific

research purposes. Potential users **must** apply to the responsible personnel for access; datasets will only be granted to users upon successful application approval.