Affective Video Summarization

by

Berkay Köprü

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ ÜNİVERSİTESİ

September 13, 2022

Affective Video Summarization

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Berkay Köprü

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Engin Erzin

_____

Prof. Yücel Yemez

_____

Prof. Albert Ali Salah

_____

Assoc. Prof. Erkut Erdem

_____

Assist. Prof. Zafer Dogan

Date: _____

I dedicate this thesis to my wife, Aysun Gurur Önalan Köprü

ABSTRACT

Affective Video Summarization
Berkay Köprü
Doctor of Philosophy in Electrical and Electronics Engineering
September 13, 2022

With the availability of video sharing and streaming services, the media consumption behavior of users has moved from TV to the internet resulting in a surge in video content. The video summarization field produces solutions for efficient video representation, retrieval, and browsing to ease the complications caused by video content and traffic surge. Inspection of the new video content shows that user-generated human-centric video production and consumption consolidates most of the surge.

Conventional video summarization neglects the human content and treats all video categories similarly. In this thesis, we argue that summarization of human-centric videos requires both understanding human behavior and video summarization. We break down this complex task into serial sub-tasks to understand complex human behavior through emotion recognition and video summarization.

First, we represent the human video content by affective states and propose a multi-modal, multi-task learning-based framework estimating affective states from audio-visual input. Along with predicting affective states, we define a novel problem of detecting affective bursts to capture salient regions in the affective contour accurately. We define affective video summarization which focuses on the summarization of human-centric videos and proposes a framework that integrates affective information into the video summarization process. Finally, we presented a dataset referred to as AffWild2-VS annotated for video summarization, enabling joint research on video summarization and emotion recognition.

ÖZETÇE

Duygu Bazlı Video Özetleme
Berkay Köprü
Elektrik ve Elektronik Mühendisliği, Doktora
13 Eylül 2022

Video paylaşımı ve yayın hizmetlerinin kullanılabilirliği ile, kullanıcıların medya tüketim davranışı TV'den internete taşındı ve bu durum video içeriğinde bir artışa neden oldu. Video özetleme alanı, video içeriği ve trafigindeki hızlı artışın neden olduğu komplikasyonları hafifletmek için verimli video gösterimi, alma ve tarama için çözümler üretir. Yeni video içeriğinin incelenmesi, kullanıcı tarafından oluşturulan insan merkezli video üretimi ve tüketiminin bu artışın çoğunu konsolide ettiğini gösteriyor.

Geleneksel video özetleme, insan içeriğini ihmal eder ve tüm video kategorilerini benzer şekilde ele alır. Bu tezde, insan merkezli videoların özetlenmesinin hem insan davranışını anlamayı hem de video özetlemeyi gerektirdiğini savunuyoruz. Duygu tanıma ve video özetleme yoluyla karmaşık insan davranışlarını anlamak için bu karmaşık görevi seri alt görevlere ayırıyoruz.

İlk olarak, insan video içeriğini duyuşsal durumlarla temsil ediyoruz ve görsel-işitsel girdiden duyuşsal durumları tahmin eden çok modlu, çok görevli, öğrenmeye dayalı bir çerçeve öneriyoruz. Afektif durumları tahmin etmenin yanı sıra, afektif konturdaki göze çarpan bölgeleri doğru bir şekilde yakalamak için afektif patlamaları tespit etmeye yönelik yeni bir problem tanımlıyoruz. İnsan merkezli videoların özetlenmesine odaklanan ve duygusal bilgileri video özetleme sürecine entegre eden bir çerçeve öneren duygusal video özetlemeyi tanımlıyoruz. Son olarak, video özetleme ve duygu tanıma üzerine ortak araştırma yapılmasını sağlayan, video özetleme için etiketlenmiş AffWild2-VS olarak adlandırılan bir veri kümesi sunduk.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Prof. Engin Erzin who granted me the opportunity to follow a Ph.D. degree and supported me through my studies. His supportive, patient, and understanding attitude eased my life as Ph.D. candidate who works in parallel. This work could not have been possible without his constant support and guidance over the past four years. His guidance led me to grow as both a researcher and as a human, and worth my life long gratitude.

I would like to thank all the administrate staff of the Graduate School of Sciences and Engineering, in particular, Emine Buyukdurmus and Hilal Demirtas. In addition, I would like to thank my jury members for their valuable time and feedback.

Finally, I dedicate this thesis to my wife, Aysun Gurur Önalan Köprü. In all honesty, without her constant support, both spiritual and technical, I would not have been able to conclude my studies. Just as a cursor for her help, I remember that she tutor me for two weeks to prepare for my PhD Qualification exam. Her belief in me was the only the thing that helped me to continue my studies when I got desperate due to rejections or setbacks. The value and love of my life, this one is for you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The internet infrastructure and services have recently matured enough for any user to easily record and share unedited videos, resulting in over 500 hours of footage being uploaded per minute to video streaming platforms like YouTube, TikTok, Twitch, Netflix, and Disney+. On top of these recent advancements, an unexpected event, the COVID-19 breakout, has caused a significant surge in internet usage[1],[2], mostly funneled to a specific section of the market entertainment. During the COVID-19 outbreak, 1-2 hours of additional daily time spent on social media platforms per user[3] has led to significant user acquisition growths in video streaming platforms. For instance, TikTok has quadrupled the monthly average hours spent per user by reaching 25 hours[4] while increasing quarterly users to 10 times from 2018 to 2021[5].

Such a rapid increase in available video content and user engagement with video platforms require automation in any video-related task such as creation, representation, and repurposing. The need for automation has motivated the research in efficient and smart Human-Computer Interaction (HCI) tools for video content. The automatic compact representation of the videos, which is the main focus of this thesis, is addressed by the automatic video summarization field.

---

[1]PCMag: Data Usage Has Increased 47 Percent During COVID-19 Quarantine

[2]Statista: Share of social media users in the United States who believe they will use YouTube more if confined at home due to the coronavirus as of March 2020

[3]Additional daily time spent on social media platforms by users in the United States due to coronavirus pandemic as of March 2020

[4]TikTok Statistics – Updated July 2022

[5]TikTok Revenue and Usage Statistics (2022)

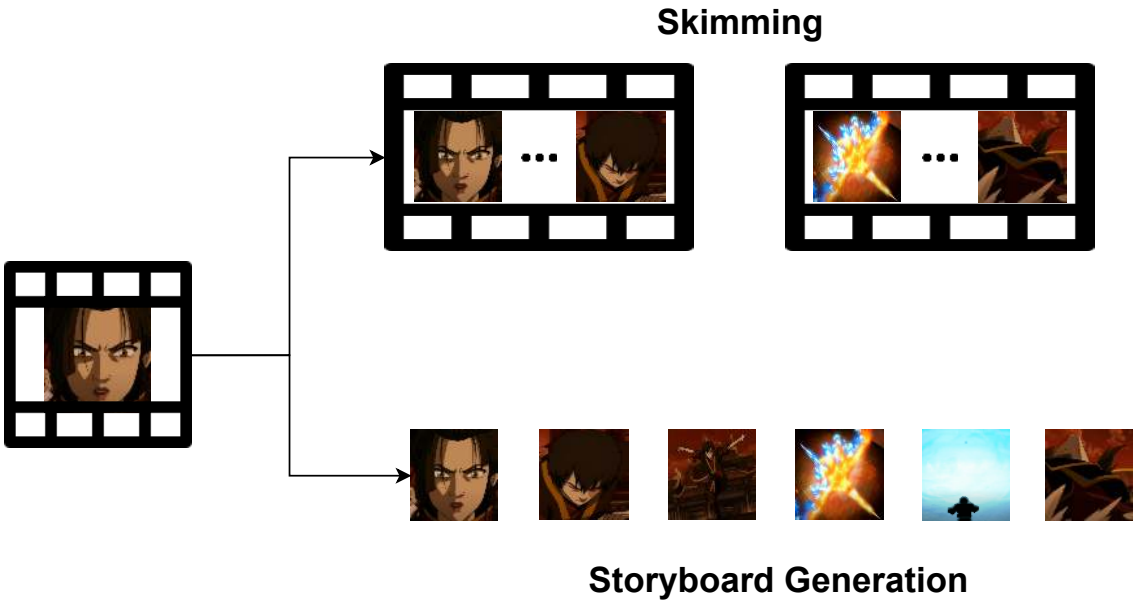


Figure 1.1: Video Summarization types

1.1 Video Summarization

Video summarization techniques can extract compact and efficient representations of videos by selecting subsets of frames from the videos. Video summarization is categorized regarding the output type which could be video fragments or video frames as depicted in Figıre 1.1[Apostolidis et al., 2021, Zhang et al., 2016, Ma et al., 2002, Ma et al., 2005]. The former, chronologically stitched segment selection, is referred to as video skimming. The major benefit of video skimming is that the summary includes both audio and motion elements, which enrich the emotions and the amount of information delivered by the summary. The latter, finding keyframes, is known as video storyboard generation. The storyboard generation lacks smoothness and naturalness due to its discrete nature. However, since it does not require synchronization, it offers more degrees of freedom in data organization than video skimming [Apostolidis et al., 2021]. This thesis focus on video summarizing techniques under the storyboard generation category.

The rise of video platforms like TikTok and YouTube has changed the video

ecosystem. The vast majority of video uploads become user-generated human-centric videos focusing on human activities, interactions, and their emotional and physical attributes [Vicol et al., 2018]. Summarization of the human-centric videos, therefore, should address the problem of selecting salient video segments or keyframes of the human-centric actions, which are frequently associated with emotional changes, from the video to construct an informative and interesting summary. However, the conventional video summarization methodologies neglect the change in particular emotional attributes, thus, are insufficient to represent the human-centric videos.

Using emotional attributes for human-centric video summarization involves modeling, quantifying, and recognizing emotions via audio-visual cues, which is studied under the affective computing domain in the literature.

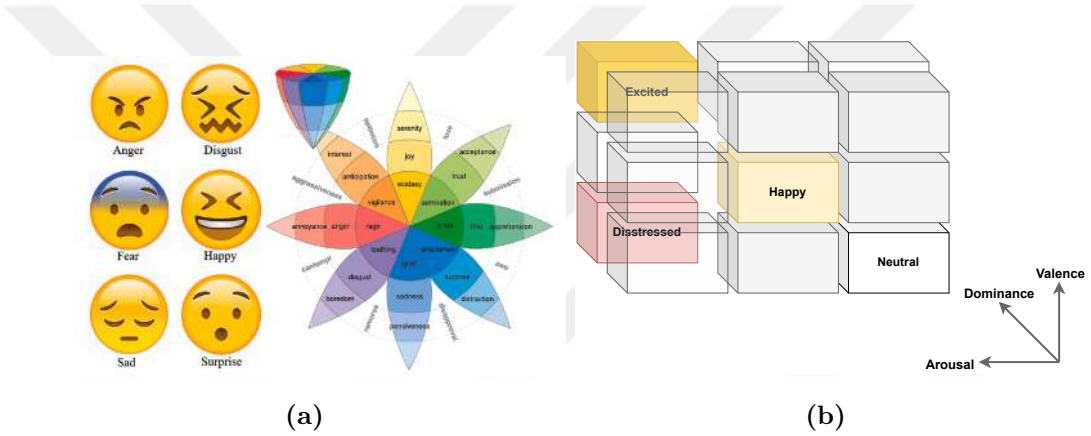
1.2 Affective Computing

Conventionally affective computing covers emotion recognition and sentiment analysis [Wang et al., 2022]. In this thesis, we only consider emotion recognition; hence we refer affective computing as emotion recognition. The emotion recognition field aims to build human-like machines capable of perceiving emotions from speech, video, sensor data, etc., and promptly making sensitive and friendly responses.

Human beings excite emotions in response to external and internal events of major significance to them. To understand and quantify these excitations, psychologists developed two typical theories to model human emotion: the discrete emotion model (or categorical emotion model) [Paul, 1999] and the dimensional emotion model [Mehrabian, 1996, Schlosberg, 1954, Russell and Mehrabian, 1977]. The dimensional model is represented by a 3-dimensional continuous space and emotions like happiness and sadness correspond to regions in this space as depicted in Figure 1.2a,. The coordinates are Activation, Valence, and Dominance (AVD), indicating activeness-passiveness, positiveness-negativeness, and dominance-submissiveness, respectively [Schlosberg, 1954, Russell and Mehrabian, 1977].

In addition, similar to [Zlatintsi et al., 2017] it is possible to distinguish three emotions in a HCI scenario: intended, expected and experienced. The intended

emotion annotations are meant to capture the emotional response that the movie or a video tries to evoke in the viewer. Intended emotions can be estimated using signals extracted from the video such as actions or speech of the human in the movie or video. On the other hand, experienced emotions represent the actual emotional experience of an individual while watching a movie or video and can only be estimated using the signals of the viewer. Finally, expected emotion is the normative emotional response, the expected emotional experience of a random viewer.



(a) **(b)**

Figure 1.2: Emotional models: (a) the Categorical emotion model, (b) Dimensional emotion model.

Along with the representations of emotions, emotions are episodes of coordinate changes in several components, including neuro-physiologic activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes [Anagnostopoulos et al., 2015], recently intensity of emotions has attracted affective computing researchers. Emotional intensity changes are studied under different terminologies such as *'affective burst detection'* and *'vocal burst detection'*. This thesis focuses on tracking dimensional attributes of emotion and affective burst detection. First, we will introduce affective state tracking and affective burst detection problems.

1.2.1 Affective State Tracking

Affective State Tracking follows the dimensional model to represent emotions and aims to predict continuous AVD attributes. Hence, it is referred to as Continuous Emotion Recognition (CER). Following the fact that humans convey their emotions through motion, facial expressions, voice, and language [Mehrabian, 2017], CER process one or multiple of these means to predict AVD.

CER research initially concentrated on producing speech-based traditional uni-modal regressors which operate on hand-crafted features from speech. As deep learning has revolutionized computer vision, speech recognition, and its applications, the trend shifted from hand-crafted feature extractions to fully deep learning-based solutions. In addition, with the availability of multi-modal databases such as CreativeIT [Metallinou et al., 2010] and RECOLA [Ringeval et al., 2013], research trend shifted to multi-modal studies where both information fusion from different modalities to enhance AVD regression. In this thesis, we develop deep-learning based on uni-modal and multi-modal CER modules by following the trend.

1.2.2 Affective Burst Detection

Affective bursts are defined in this thesis as sections in the affective contour where affective attributes change significantly. Hence, this definition contains non-vocal expressions such as laughing and screaming and vocal expressions such as accentuations.

Affective burst detection enables researchers to model inter-emotion changes (transition from happiness to excitement) and intra-emotion changes. In addition, detecting salient region at the affective contour with affective burst could help salient section/region or frame detection in the speech, images, or video.

1.3 Affective Video Summarization

While 'human-centric videos' are taking over most of the video content, the tremendous progress in the VS task has not quite developed for both 'human-centric video

understanding' and 'human-centric video representation.' 'Human-centric video representation' involves multiple downstream task: *Continuous Emotion Recognition* and *Video Summarization*. However, annotated data for emotional modeling and video representation is lacking. As a result of the lack of data, there is a lack of interest.

We aim to explore affective information as an additional modality to improve the summarization of human-centric videos rather than detecting salient affective regions via video summarization techniques. In doing so, we model the intended emotions and emotional processes that are contained in the video rather than experienced emotions.

In this thesis, we address the use of affective content in human-centric video summarization to meet the demand for efficient representation of this new plethora of video content and refer to it as Affective Video Summarization (AVS).

1.4 Problem Statement

AVS is a complex task, and it requires a solution to both CER and VS. In the following, the dealt problems are listed sequentially:

Research Question 1: How to represent and extract salient affective content from audio and video?

Representation of affective content is still an ongoing study for HCI applications. As understanding and recognition of emotion is a such natural behavior for human-beings, it is not straightforward what is the best way of representing affective content to drive VS.

Assuming there is a universal and effective representation of affective content for VS, the question arises: What is the reliable way to extract this representation?

Chapter 2 will explore the relationship between affective attributes (AVD) and their behavior in the time domain. Following these relationships, we will explore multiple representations of affective content. Further, we will study multi-modality (audio-visual) and uni-modality (audio) to extract these representations.

Research Question 2: How to use extracted affective content to drive

video-summarization for human-centric videos?

The second step for summarizing human-centric videos is formulating AVS problem and exploring effective ways to inject affective representations into the conventional video summarization frameworks.

Chapter 3 will explore state-of-art datasets for video summarization for human-centric content and formulate AVS. Following the AVS formulation, information fusion techniques for visual and affective content will be explored.

Research Question 3: How to jointly optimize affective content extraction and video-summarization for human-centric videos?

The lack of datasets for AVS including both annotations for affective states and video summarization, forces a joint optimization problem to be solved sub-optimally and sequentially.

Chapter 4 will introduce a novel dataset for AVS, AffWild2-VS, whose video summarization annotations are collected on top of affective annotations.

1.5 Contributions of this Thesis

The main contribution of this thesis revolves around the summarization of human-centric videos, and to achieve our goal, we have also provided minor contributions to downstream fields.

Major Contributions:

- We formulate a novel end-to-end learning problem for affective-information enriched video summarization targeting human-centric videos.
- We model affective information in terms of emotional attributes and learned embeddings extracted from the CER module.
- We investigate the use of attention mechanisms in video summarization for human-centric videos and develop temporal and spatial attention-based frameworks.

- We investigate conventional video summarization evaluation frameworks and introduce a novel evaluation metric targeting evaluation of human-centric video summarization.

- We collect video summarization annotations of famous affective state tracking dataset, enabling joint optimization of CER and VS.

Minor Contributions:

- We study multi-modality for CER, and develop multi-modal deep neural network architecture based on speech and body motion.

- We study correlations between AVD attributes and make use of their correlations to build multi-task learning based multi-head deep neural network to estimate AVD jointly.

- We introduce a novel correlation loss to train CER estimator neural networks.

- We formulate a novel affective burst detection problem capturing the high-intensity changes, which can improve the understanding of the inter-emotion and intra-emotion transitions in the scene.

1.6 Thesis Structure

The remainder of this thesis is organized into four chapters. In chapter 2, we address AVD estimation and ABD. First, we explore the correlations between AVD attributes and propose a multi-modal multi-task convolutional recurrent network. Then, noting that the CER modules fail to capture high-intensity changes, we define the ABD problem and propose an audio-based convolutional neural network to detect AFFB. In chapter 3, we address AVS problem. First, we explore the state-of-art datasets for human content. Then propose a novel affect-driven video summarization based on convolutional neural networks and attention. In chapter 4, we stress the need for a dataset-specific for AVS. Then, we introduce AffWild2-VS dataset

having the annotations for CER and VS. Finally, in chapter 5, we summarize the thesis, and discuss future directions.

Chapter 2

AFFECTIVE COMPUTING

Human beings excite emotions in response to external and internal events of major significance to itself. The emission of emotions, also known as excitations, are represented by a 3-dimensional continuous space, and emotions like happiness and sadness correponds to regions in this space. The coordinates are Arousal, Valence, and Dominance (AVD), indicating activeness-passiveness, positiveness-negativeness, and dominance-submissiveness respectively [Schlosberg, 1954, Russell and Mehrabian, 1977]. Emotions are episodes of coordinate changes in several components, including neuro-physiologic activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes [Anagnostopoulos et al., 2015].

In the context of emotion recognition, the continuity of the labels separates into Continuous Emotion Recognition (CER) and Emotion Classification (EC). CER research focuses on prediction of AVD values [Fatima and Erzin, 2017, Tzirakis et al., 2017], while EC focuses on classification of discrete emotions [Castellano, 2008, Castellano et al., 2008, Ng et al., 2015, Mirsamadi et al., 2017]. Another significant difference between CER and EC studies is the loss function that is used during training. While CER studies are mandated by mean squared error (MSE) [Fatima and Erzin, 2017], EC is mandated by cross-entropy [Mirsamadi et al., 2017].

CER research initially concentrated on the speech, and produced uni-modal regressors as speech not only highly reflects the neuro-physiological changes in the research but also its maturation in terms of analysis [Nicholson et al., 1999, Kwon et al., 2003, Schmitt et al., 2019]. However, studies were recently directed into multimodality where descriptors from body, face, speech and text are combined [Fatima and Erzin, 2017, Chen et al., 2017]. Especially, combining speech with body motion

and facial expressions has become a common practice, and multi-modal approaches have quickly proven their superiority against the uni-modal approaches [Fatima and Erzin, 2017, Chen et al., 2017].

Traditionally, CER set-up consists of a feature extraction, feature summarization and regression blocks [Fatima and Erzin, 2017, Han et al., 2014, Castellano, 2008]. With the rise of Deep Neural Networks (DNN) and their ability to model non-linear behaviors, initially the Feed-Forward Networks occupied the regressor blocks [Han et al., 2014]. Later, Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) were introduced into CER. RNNs replaced the Feed-Forward Networks to capture long term relationships [Han et al., 2017], while CNNs were adapted successfully as a meta-feature extractor by [Khorram et al., 2017, Castellano et al., 2008, Khorram et al., 2018]. As extracting low level descriptors such as acoustic features requires domain expertise and suffers from task independence, end-to-end learning, where raw signals are fed into deep architectures, were proposed as the replacement [Graves and Jaitly, 2014]. Such end-to-end systems then successfully adapted to the emotion recognition problem by [Trigeorgis et al., 2016, Tzirakis et al., 2017, Zhang et al., 2019].

The state-of-art approaches attack the CER problem by training a predictor for each output. This approach not only increases the complexity of the general task but also suffers from over-fitting due to limited amount of training data. By training the predictor with multi-task learning (MTL), one can force it to learn shared representations and increase the generalization ability [Baxter, 1997]. Although MTL is exploited by [Zhang et al., 2019, Parthasarathy and Busso, 2017, Chen et al., 2017] in the context of CER, unlike our approach these studies trained the proposed architectures with the MSE loss.

Studies on CER evaluate themselves regarding correlation metrics Concordance Correlation Coefficient (CCC) [Schmitt et al., 2019], or Pearson Correlation Coefficient (PCC) [Fatima and Erzin, 2017] while training the predictors using different type of error metrics such as MSE or mean absolute error (MAE) loss . This mismatch causes a degradation in the performance of proposed architectures as also

discussed in [Trigeorgis et al., 2016, Tzirakis et al., 2017].

In CER studies, correlation-based metrics are widely used for evaluation tasks since the trend of the predicted attribute is more important than the actual level of the prediction. On the other hand, correlation metrics do not always grant effective capturing of all high-intensity changes in the affective domain.

Due to the categorized nature of the DER, changes between the categories (inter-emotion transitions) can be observed, however transitions within each category (intra-emotion transitions) do not appear or are not typically available for the DER studies. On the other hand, continuous affect attributes in the CER problem provide the necessary intensity fluctuations for the inter-emotion and intra-emotion transitions.

The detection of inter- and intra-emotion transitions, i.e., affective change detection, has a significant importance, and it is widely studied in the psychology domain under the *mismatch negativity* (MMN) literature [Male et al., 2020, Kovarski et al., 2017, Kojouharova et al., 2019].

Affective change detection is studied in the DER context by [Huang et al., 2015] and CER context by [Parthasarathy and Busso, 2018]. *Affective change points* are defined as the transition points between the emotions in [Huang et al., 2015]. These points are estimated using a Gaussian mixture model-based architecture with and without prior emotion class information. In [Parthasarathy and Busso, 2018], emotional hotspots are defined as sections deviating from the affective attribute's median. They proposed a qualitative agreement-based assessment method to map affective attributes into low, high, neutral, and non-consensus sections. A section is labeled as low if that section is under the median and high if it is above the median. Then, bidirectional long short-term memory (BLSTM) is operated on 88 eGeMAPS [Eyben et al., 2016] features extracted from acoustic signals to classify the trend. However, this approach highlights flat sections that deviate from the median. It resembles a quantization approach more than tracking the high-intensity regions as the labeling procedure misses all the inter- and some intra-emotion transition regions in the affective domain.

In this study, first, we address the discrepancy between CER's evaluation and training process by defining a loss aligned with evaluation metrics. In addition, to utilize the correlations between AVD, we introduce CER-MTL that jointly estimates AVD attributes. While our CER model follows the general trend successfully, it fails to detect high-intensity changes by nature; thus, in addition to CER, we address the segmentation of the affect contour into affective burst and idle regions. Unlike [Parthasarathy and Busso, 2018], we label sections regarding the intensity of change rather than their difference to the median value. Affective burst sections were then estimated using spectral features of speech as input and a kernel-fusion dilated convolutional neural network (KFDCNN) as the classier. To summarize, the main contributions of Chapter 2 are as follows:

- A multi-modal deep neural network architecture based on speech and body motion (CER-MTL) is proposed
- A MTL-based architecture surpassing single-task learning (STL) based state-of-art solutions is proposed
- A correlation loss metric that is used as an objective function during the training is introduced
- A new labeling mechanism to define the affect contour's high intensity and idle segments is proposed.
- We formulate a novel affective burst detection problem capturing the high-intensity changes, which can improve the understanding of the inter-emotion and intra-emotion transitions in the scene.
- We propose a novel architecture KFDCNN for affective burst detection from speech.
- We evaluate the RECOLA and CreativeIT datasets with classification metrics.

The rest of Chapter 2 is structured as in the following. First, we will present our framework for CER in Chapter 2.1, and provide experimental evaluations of CER-MTL on RECOLA and CreativeIT datasets. Then, we will introduce the affective burst detection problem and our solution KFDCNN in Chapter 2.2. Finally, we will discuss our findings on CER and affective burst detection in Chapter 2.3.

2.1 Multi-modal Continuous Emotion Recognition using Multi-Task Learning

In this section, we first define the feature representations of the speech and body motion signals. Then, the dataset that the experiments are conducted on is introduced, and finally, we provide the details of the proposed framework.

2.1.1 Feature Representations

Mel-frequency cepstral coefficients (MFCC) are extracted from the speech signals, and utilized to generate the acoustic features. Each speech frame is summarized into the 39 dimensional acoustic feature vector which includes the energy, the first 12 MFCCs and their first and second derivatives. The MFCC feature vector is represented as $f_l^s \in \mathbb{R}^{39\times 1}$ for the lth time frame.

To represent the body motion, at each instance raw Euler rotation angles in (x,y,z) dimensions along with their deltas are utilized. The resulting body motion feature vector for the lth frame is defined as $f_l^b \in \mathbb{R}^{24\times 1}$.

2.1.2 Dataset

In this study, USC CreativeIT [Metallinou et al., 2010] and RECOLA databases are used to train the proposed architecture. The CreativeIT database consists of pairs of 3.5 minutes length interplays performed by 16 actors to mimic dyadic interactions. After removing recordings with missing labels for any AVD, the training dataset is generated from 40 recordings. Then the recordings are divided into five sessions regarding the mutual exclusiveness of the speakers. Dividing the dataset into these mutually exclusive sessions enables speaker-independent CER.

The RECOLA database is a popular database for audiovisual emotion recognition, as it is used for the AVEC challenge [Ringeval et al., 2015],[Valstar et al., 2016], and [Ringeval et al., 2018]. The database contains dyadic interactions of 27 French-speaking subjects. The recordings are separated into three parts 9 train subjects, 9 development, and 9 test subjects, as the AVEC challenges urge. By following the

procedure in [Wu et al., 2019], only the train set is used for training, and results are provided from the development set. The database is annotated with frame rate of 40 ms by six annotators. In this study, the mean of these 6 annotations are used as the reference. In addition, while experimenting on RECOLA, body Euler angles in Figure 2.1, are replaced with the facial activation units, yaw, pitch, mean and standard deviation of optical flow.

2.1.3 Framing

To capture the temporal continuity and variations in the features and to consider the slow varying nature of emotions, we choose a sequence of feature vectors to form a temporal feature-image representation. Since the frame rate of the speech and body features is 60 fps, we choose to form the sequence of feature vectors with a stride of $t = 10$ frames as

$$F_l = [f_{l+1-Nt/2}, \ldots, f_{l-t}, f_l, f_{l+t}, f_{l+2t}, \ldots, f_{l+Nt/2}] \tag{2.1}$$

where $F_l \in \mathbb{R}^{P \times N}$ is the feature image at frame l, P is the feature vector dimension and N is the number of frames defining the temporal extent.

2.1.4 Architecture

The proposed semi-end-to-end architecture is depicted in Figure 2.1. The proposed architecture mimics a single-task end-to-end framework which is composed of multiple CNN layers, Max-Pooling Layers, RNN layers and a final fully connected layer [Trigeorgis et al., 2016]. In this architecture, CNN acts as a meta-feature extractor that can extract task-dependent features or meta-features from the raw signal or input features.

RNN's are famous for their ability to model sequential information. A combination of CNN and RNN layers are effectively used to find the long-term relationships for the image captioning task [Vinyals et al., 2015, Xu et al., 2015].

On top of the single end-to-end framework (CER-STL), the proposed architecture involves MTL, and these tasks share the same meta-feature extractor. We will refer

to the proposed architecture as CER-MTL.

Modality Fusion

Our framework applies an early fusion by concatenating the modalities at the input. Hence, the input of the multimodal architecture is $F_l^{b,s} \in \mathbb{R}^{63\times N}$. Concatenating the modalities at the input is both beneficial in terms of complexity of the total architecture, and also from the learning point of view. The most complexity-hungry part of the proposed architecture is the CNN layer due to the large input dimensionality. As a result, using the same layer for each task benefits from complexity. In addition, early fusion enables the learning of cross-modal features.

Loss

In this study, we utilize the PCC and CCC as loss functions, in mutually exclusive setups, between the batch of ground truth and predictions.

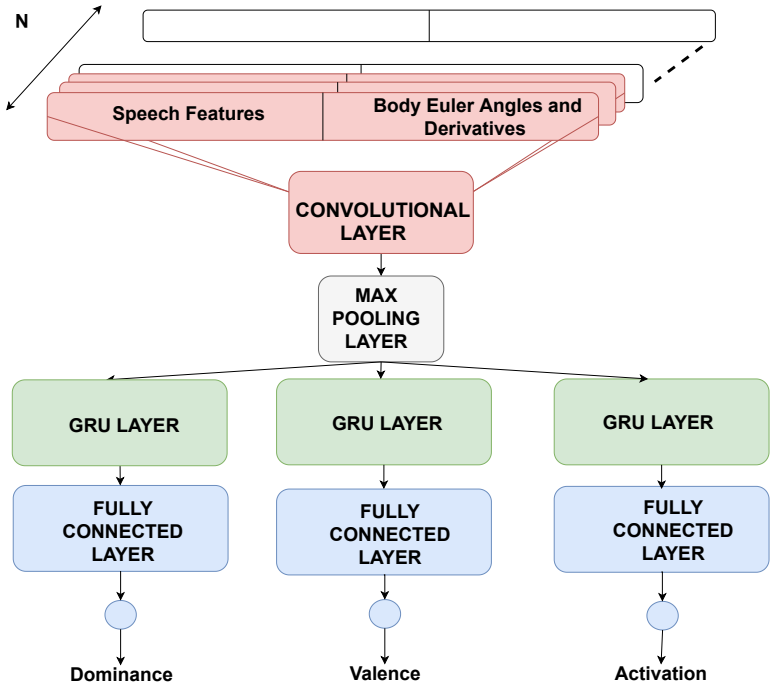


Figure 2.1: Semi-End-to-End framework exploiting MTL for CER

Let the $\hat{y}$ be the output of CER-MTL, and y be the corresponding ground truth vector then the PCC (L^{PCC}) loss is defined as:

$$\rho_{PCC} = \frac{\mathbf{E}[(y - \mu_y)(\hat{y} - \mu_{\hat{y}})]}{\sigma_y \sigma_{hat_y}}, \tag{2.2}$$

$$L^{PCC} = 1 - \rho_{PCC}. \tag{2.3}$$

where ρ_{PCC} is PCC. Similarly, CCC loss (L^{CCC}) is defined as:

$$\rho_{CCC} = \frac{2\rho_{PCC}\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{2.4}$$

$$L^{CCC} = 1 - \rho_{CCC}, \tag{2.5}$$

where ρ_{CCC} is CCC. MSE Loss (L^{MSE}) is defined as:

$$L^{MSE} = \mathbf{E}[(y - \hat{y})^2]. \tag{2.6}$$

Multiple Task Learning

The objective function of the MTL could be written as the weighted sum of objective functions of individual tasks which are predicting arousal, valence, or dominance. Let $J(\theta)$ be the objective function of MTL given as

$$J(\theta) = \sum_{m=1}^{M} \alpha_m L_m, \tag{2.7}$$

where α_m is the weight of the loss of the m^{th} task L_m where $m \in \{\text{arousal}, \text{valence}, \text{dominance}\}$.

2.1.5 Experimental Evaluation

The proposed architecture is implemented in Keras framework with TensorFlow backend. All the conducted experiments are held on NVIDIA GeForce GTx 1080 Ti. During the training Adam optimizer with a learning rate of 0.00005 is used. The training is held with a batch size of 256 for 100 epochs with an early stopping option of 20 epoch. All these hyper-parameters are optimized during the experiments.

The CNN layer of the CER networks has 25 filters with kernel size of 3, and the max-pooling layer has a kernel size of 2. The output features are then fed into 5 GRU nodes per task. To add further non-linearity the GRU's outputs are fed into a fully connected layer having 2 node, and finally a linear node is added to regress the AVD values. The weights of the MTL are selected uniform as $\alpha_m = 1/3$ for $m \in \{1, 2, 3\}$ for CreativeIT and $\alpha_m = 1/2$ for $m \in \{1, 2\}$ for RECOLA to not prioritize a task over another.

To depict the effectiveness of the proposed architecture we devised experiments comparing MTL with STL, correlation losses (PCC Equation 2.3 and CCC Equation 2.5) with MSE loss Equation 2.6, multi-modality with single modality and different window sizes. For all the experiments held, CCC is used as the evaluation metric.

Table 2.1: Mean CCC comparison between CCC loss, PCC loss and MSE loss on CreativeIT database for unimodal and multimodal settings ($N = 20$, temporal extent is $Nt = 3.3$ sec)

Model	**CCC Loss**			**PCC Loss**			**MSE Loss**		
	Act	Val	Dom	Act	Val	Dom	Act	Val	Dom
Multimodal	0.38	0.16	0.18	0.33	0.16	0.14	0.26	0.07	0.11
Motion	0.15	0.06	0.10	0.14	0.02	0.09	0.12	0	0.04
Speech	0.37	0.13	0.18	0.32	0.13	0.18	0.23	0.08	0.06

Table 2.2: Mean CCC comparison between CCC loss, PCC loss and MSE loss on RECOLA database for unimodal and multimodal settings ($N = 20$, temporal extent is $Nt = 3.3$ sec)

Model	**CCC Loss**		**PCC Loss**		**MSE Loss**	
	Act	Val	Act	Val	Act	Val
Multimodal	0.66	0.34	0.49	0.24	0.50	0.21
Facial Attributes	0.15	0.06	0.09	0.18	0.22	0.19
Speech	0.61	0.32	0.34	0.13	0.49	0.19

Results in Table 2.1 demonstrate the effect of loss function and using different modalities for CER. For the multimodal case, CER-MTL with CCC loss outperforms the same architecture with PCC and MSE losses, while PCC loss outperforms MSE loss for all the of prediction modalities. For instance the proposed framework achieves 0.38 CCC for arousal when it is trained with CCC loss, the same architecture achieves only 0.33 and 0.26 CCC's when it is trained with respectively PCC and MSE losses. These results state that using correlation based metrics are more suitable as the loss function when the evaluation metrics is also correlation based. In fact, with the results presented in Table 2.2 where CCC loss achieves at least 10% improvements, PCC loss fails to outperform MSE loss on prediction of arousal. With regard to this observation, our argument on correlation metrics and losses can be strengthen as, using evaluation metric as the loss function of the DNN architecture provides a robust performance enhancement. However, the downfall of the correlation based loss functions are they require longer batch sizes than regular error metrics like MSE. As the longer batch sizes prone to over-fitting the learning rates should be chosen carefully. In our study, we used relatively small learning rates like between 0.0001 and 0.00005. However, our findings regarding over-fitting and learning rate are aligned with [Zhang et al., 2019].

The Figure 2.2 depicts the predicted AVD values and the CCC correlation of

the predictions from CER-MTL with CCC loss, PCC loss, and MSE loss over 256 frames, where the CCC values at time index l are calculated by considering the previous 100 predictions before time l. The proposed CCC training outperforms the MSE training by at least 10% CCC difference. In addition, MSE training is unresponsive to the significant changes in the reference, while CCC and PCC loss is highly responsive.

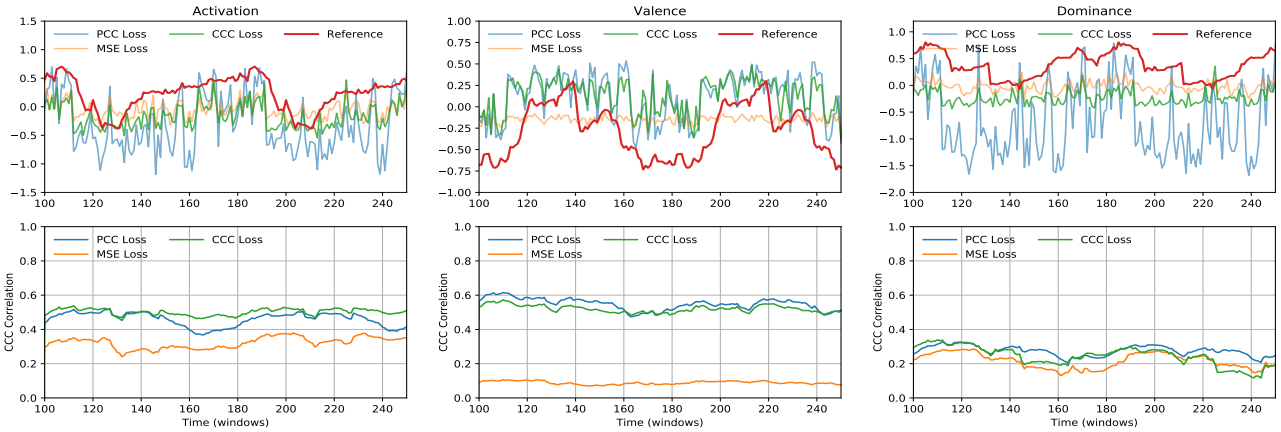


Figure 2.2: Prediction of AVD modalities over time for correlation loss and MSE loss

The performance comparison of unimodal and multimodal architectures is presented in Table 2.1. Our experiments show that speech carry more information about affective states as it outperforms the motion based architecture with at least 10% CCC difference on creative IT and 2% CCC on RECOLA. Especially, we found that body motion is almost uncorrelated with valence attribute. Another observation in regardless of the loss function arousal and valence attributes are mandated by the speech modality.

The CCC results of the proposed MTL system and the baseline STL system are given in Table 2.3. CER-MTL outperforms CER-STL on predicting arousal and valence attributes. Especially, MTL brings 6% CCC improvement for valence. From Table 2.3 one can deduce that learning features that are significant for both arousal and dominance helps predicting the valence. MTL improves the results not

Table 2.3: Mean CCC comparison between CER-MTL and CER-STL multimodal architecture CCC loss with $N = 20$ (temporal extent is $Nt = 3.3$ sec)

Model	**CreativeIT**			**RECOLA**	
	Act	Val	Dom	Act	Val
MTL	0.38	0.16	0.18	0.66	0.34
STL	0.31	0.10	0.16	0.62	0.27

only by creating a regularization effect for the other tasks, but also for the special case of CER, it exploits the correlation between the AVD values. Another benefit of the MTL arises in terms of complexity and resource efficiency. Baseline results are generated by training CER-STL for each task. Training time and the total complexity of the networks for the baseline is almost multiple of the total number of task.

Table 2.4 demonstrates the effect of using different window sizes for CER. Note that increasing the window size significantly effect the prediction of arousal while there is no such significant trend for valence and dominance. Intuitively increasing receptive field size would add more information to be learned, the slow varying nature of affective states might add repetition rather than information. Hence, the increasing the input size directly effects the complexity/number of parameters of the CER-MTL which makes it prone to over-fitting.

Table 2.5 presents the performance comparison of the proposed CER-MTL architecture with the state-of-the-art solutions on the RECOLA dataset. CER-MTL ($N = 60$) outperforms the state-of-the-art solutions by at least 1% CCC on arousal prediction. This comparison represents the power of the proposed architecture. On the other hand, CER-MTL is less competitive at predicting the valence attribute which might be due to speech features overwhelming the facial attributes as it performs similarly to the multimodal architecture introduced in [Tzirakis et al., 2017].

Table 2.4: Effect of temporal extent (N) on CER for the MTL model with the CCC loss

N	**CreativeIT**			**RECOLA**	
	Act	Val	Dom	Act	Val
20	0.38	0.16	0.18	0.66	0.34
40	0.42	0.16	0.18	0.74	0.35
60	0.45	0.16	0.18	0.79	0.37

Table 2.5: Comparison of CCC performances on the development set of RECOLA database

Model	**Act**	**Val**
Pure CNN [Schmitt et al., 2019]	0.51	0.26
End-to-end [Trigeorgis et al., 2016]	0.74	0.33
DDAT [Zhang et al., 2018]	0.78	0.50
End-to-end Multimodal [Tzirakis et al., 2017]	0.75	0.41
FER-P&G-Net [Wu et al., 2019]	0.60	0.69
CER-MTL ($N = 60$)	**0.79**	0.37

2.2 Kernel Fusion Dilated Convolutional Neural Networks for Affective Burst Detection

We propose a convolutional neural network (CNN) architecture where parallel convolutions process the input with different dilated kernel lengths to detect affective bursts. In this section, the affective burst detection problem is first defined, then feature extraction from speech signals is described. Later, an affective burst detection framework based on kernel fusion and CNN's is described.

2.2.1 Affective Bursts

Affective burst detection is a two-class classification problem where the affective contour is segmented into affective burst and idle classes. We define the affective burst segment (ABS) as the region in which the affect attribute contour is changing rapidly with high gradients. Respectively, idle segments (ISs) cover the complement of the ABSs and correspond to the regions where the affect attribute contour is changing slowly with low gradients.

The ground-truth ABS annotations are generated in two steps. First affective burst points (ABPs) on the affect attribute contours are detected, then these points are then extended into segments referred to as ABSs. Note that all non-ABS regions are referred to as idle segments.

Affective burst points (ABPs) are set based on the first-order regression coefficients of the arousal and valence attributes as

$$d_e[n] = \frac{\sum_{l=1}^{L}(e[n+l] - e[n-l])l}{2\sum_{l=1}^{L} l^2} \tag{2.8}$$

where $d_e[n]$ is the delta coefficient of attribute e (Arousal or Valence) at sample index n. By selecting a threshold τ_e, we can define the ABP indicator function p_e at sample index n as

$$p_e[n] = \begin{cases} 0 & \text{if } \tau_e < |d_e[n]| \\ 1 & \text{if } |d_e[n]| \geq \tau_e. \end{cases} \tag{2.9}$$

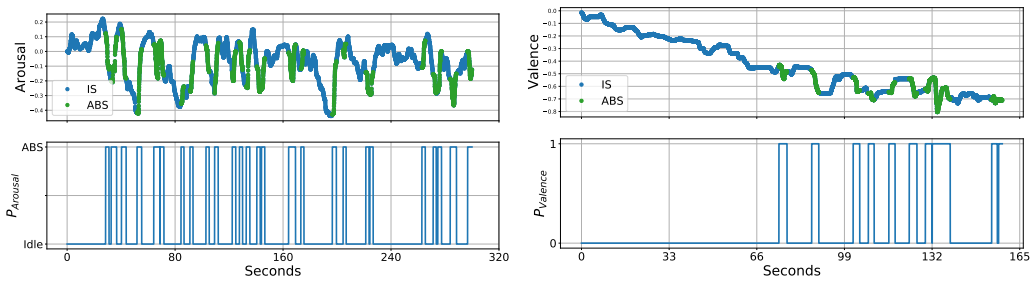
Then the ground truth binary segment labels are extracted as

$$P_e[n+i] = \begin{cases} 1 & \forall n \text{ s.t. } p_e[n] = 1 \text{ and } i = -\Delta, \ldots, \Delta \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

where an ABS of temporal size $w_s = 2\Delta + 1$, where Δ is the defining parameter of temporal extend in forward and backward direction, is centered for each ABP on the indicator function $p_e[n]$. Note that the resulting size of ABSs can be longer than w_s when two consecutive ABPs are closer than w_s. Sample affect attribute contours for arousal and valence, indicated as ABS and IS in different colors, together with $P_e[n]$ are depicted in Figure 2.3.

2.2.2 Feature Extraction

In this section, we use the *extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)* representing the spectral and temporal characterization of speech signal for ABS detection [Eyben et al., 2016]. The 88-dimensional eGeMAPS features are calculated as statistics of 25 *low-level descriptors* (LLDs), such as Mel-frequency Cepstral Coefficients, Pitch, and Loudness, using the OpenSMILE toolkit [Eyben et al., 2010].



(a) **(b)**

Figure 2.3: Sample Arousal and Valence contours with affective burst and idle segment labelings calculated using Equation 2.10; (a) Labeled affective burst segments (ABS) for Arousal from RECOLA dataset, (b) Labeled affective burst segments (ABS) for Valence from CreativeIT dataset.

The LLDs are extracted using a window size of 20 ms, and the 88-dimensional eGeMAPS features are calculated over 500 ms with a hop duration of 40 ms. Hence the eGeMAPS features are extracted at 25 fps and represented at time frame n as $F[n] \in \mathbf{R}^{88}$.

2.2.3 Kernel Fusion Convolutional Neural Network

The proposed KFDCNN architecture is depicted in Figure 2.4. KFDCNN comprises both dilated parallel and typical convolutional layers, max-pooling layer, and multiple fully connected layers. The kernel fusion layer at KFDCNN includes dilated convolutional kernels with different lengths that help learn temporal relations in different resolutions. Hence, this layer enriches representation for the ABS detection.

Furthermore, with dilated kernels, each layer has longer receptive fields, resulting in convolution outputs that capture long-term information, which is especially important for slowly varying emotional processes.

The proposed KFDCNN architecture processes a window of features which is represented as $I[n] = \{F[n-T], F[\ n\ -\ T\ +\ s],\ \ldots,\ F[n+T-s], F[n+T]\}$ where s is the dilation rate, $2T$ represents the receptive field size at the input, and $I[n] \in \mathbf{R}^{\frac{2T}{s} \times 88}$. Then KFDCNN outputs $\hat{y}[n] \in \mathbf{R}^2$ from the input $I[n]$.

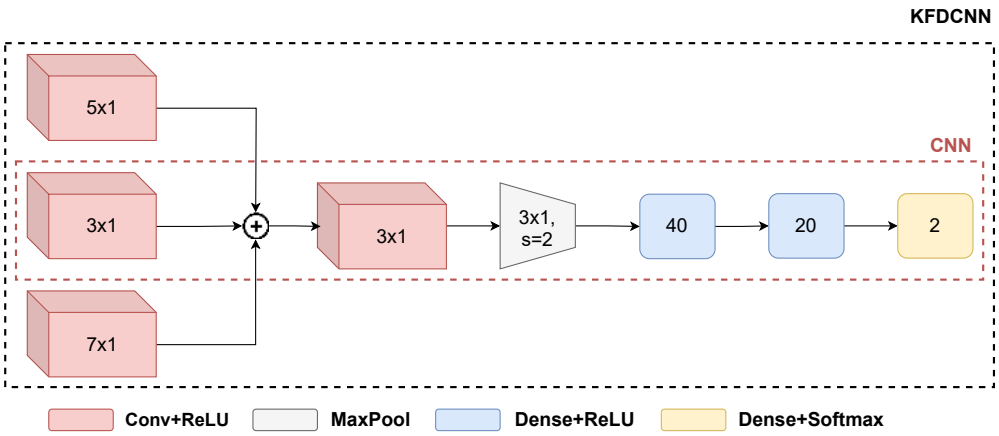


Figure 2.4: Kernel fusion dilated convolutional neural network for affective burst detection, where the convolutional layers at the input have a dilation rate of 5 **(see Chapter 2.2.3 for explanation)**

2.2.4 Model Training

The architecture is trained for ABS detection through a binary classification task. As the training scheme is the same for both attributes, for the sake of simplicity, we will drop the attribute indicator e from the P and other variables that are introduced in Equation 2.11. However, this task has an imbalanced nature. There are a small number of sections labeled as ABS, while a high number of sections are labeled as IS. To overcome the imbalance problem, we adopted weighted negative log-likelihood ratio loss:

$$L(\mathbf{P}, \hat{\mathbf{y}}, \theta) = \frac{1}{(\theta_0 + \theta_1)} \sum_n (-\theta_{P[n]} \log(\hat{y}_{P[n]}[n]])) \tag{2.11}$$

where θ_i is the weight for class i, $P[n] \in \{0, 1\}$ is the binary segment label at time frame n, $\hat{y}_{P[n]}[n]$ is the probability output of KFDCNN corresponding to the segment label $P[n]$ at frame n, where $\hat{y}[n] = \{\hat{y}_0[n], \hat{y}_1[n]\}$. The class weight θ_i is extracted as

$$\theta_i = \frac{1}{\text{frequency of } i^{th} \text{ class}} \tag{2.12}$$

for class index $i = 0, 1$ corresponding to the IS and ABS.

2.2.5 Affective Burst Detection Experimental Evaluations

The proposed architecture in Section 2.2 is evaluated on the RECOLA [Ringeval et al., 2013] and the CreativeIT [Metallinou et al., 2010] datasets. In this section, datasets and implementation details are introduced. Then evaluation metrics are described. Finally, the performance of the KFDCNN is compared against a baseline feed-forward neural network (FFN) together with the CNN and the dilated CNN (DCNN).

Datasets

We train and evaluate the ABS detection task on the widely used multi-modal datasets RECOLA and CreativeIT that are introduced in Chapter 2.1.2.

USC CreativeIT is a multi-modal database of theatrical improvisations. It is divided into five mutually exclusive sessions in terms of speakers. The cross-validation procedure on this dataset is held by leave-one session out to preserve speaker independence.

Table 2.6: Statistics over the RECOLA and CreativeIT datasets after the ground-truth ABS annotations on Arousal (A) and Valence (V) contours: Number of ABSs, mean duration of ABSs, total duration of ABSs, and mean absolute delta ($|d|$) of the ABSs

Stats	**RECOLA**		**CreativeIT**	
	A	V	A	V
# ABSs	446	461	641	632
Mean ABS dur (sec)	3.4	3.8	3.6	3.6
Total ABS dur (sec)	1510	1744	2308	2296
Mean $\|d\|$ of ABSs	0.0030	0.0017	0.0014	0.0009
Total dur (sec)	5400	5400	7708	7708

Table 2.6 presents the statistical characterization of the ground-truth ABS an-

notations on the arousal and valence contours of the RECOLA and CreativeIT datasets. Total ABS region durations cover around 30% of the datasets with similar mean ABS durations of 3.6 seconds. Mean absolute delta ($|d|$) values for the ABS regions are observed higher for the RECOLA dataset, indicating higher affect contour changes for the RECOLA.

Implementation Details and Setup

Experimental evaluations of the proposed ABS detection system are executed using cross-validation. For the RECOLA dataset, two videos are selected as the test set, and the rest are chosen for training and validation, resulting in 9 folds. On the other hand, we apply a leave-one-session-out strategy for the CreativeIT dataset, for each fold, a session is chosen as a test set, and the rest are used for training and validation, resulting in 5 folds.

We set the length, L, of the first-order regression coefficients to capture 0.8 seconds temporally, the ABS temporal window size, w_s, is set to span 2 seconds. In (2.9), we set two thresholds, τ, values, one for each dataset, so to cover 30% of the datasets as the ABS regions.

The input of the KFDCNN, $I[n]$, is set with $T = 100$ and $s = 5$ which spans an 8 seconds temporal window with dilation rate 5. Kernel fusion layer at KFDCNN has three parallel 1-dimensional convolutions with kernels sizes 3, 5, and 7, and these kernels have a dilation rate of $s = 5$. The second 1-dimensional convolution has a kernel length of 3 with a dilation rate of 1. The max-pool layer down-samples the temporal dimension in half. The output of the max-pool layer is flattened from 2-dimension into 1-dimension and fed into fully connected layers with node sizes of 40, 20, and 2, respectively.

Single kernel and no dilation derivatives of the KFDCNN are also defined and evaluated to assess the proposed model's performance better. The CNN architecture, which is depicted within the red dashed lines in Figure 2.4, has a single kernel set with a size of 3 and a dilation rate of 1. In order to have comparable complexity with the KFDCNN, the input feature of the CNN is set with $T = 20$ and $s = 1$,

which spans a 1.6 seconds long temporal window without dilation. A dilated CNN (DCNN) architecture is also defined by setting the input feature representation with $T = 100$ and $s = 5$.

Affective Change Point Detection Performances

Unweighted average F-score (UAF1) and Recall (UAR) metrics are computed at frame level via cross-validation and used for the performance evaluations.

Table 2.7 presents F1-score and Recall performances of the KFDCNN against the baselines (SVM and FFN), CNN, and DCNN. CNN-based architectures outperform the baseline in all comparison metrics by at least 2 percent-point (pp) at RECOLA and CreativeIT databases. This result stresses the importance of temporal information for the detection of ABSs. KFDCNN distinctly performs better than CNN and DCNN models among the CNN-based architectures. Performance improvement for the KFDCNN is most significant for arousal and valence in the RECOLA dataset, while only for valence in the CreativeIT dataset.

Comparing CNN with DCNN, dilation improves the F-score performance by 5 pp for arousal and 3 pp for valence in the RECOLA dataset. Moreover, similar improvements are also seen with the CreativeIT dataset. Significant temporal context due to dilated kernels is crucial to differentiate idle sections from ABS. This observation supports that consecutive temporal features carry less extrinsic information due to the slowly varying nature of emotional processes.

Comparing DCNN with KFDCNN, kernel fusion improves F-score approximately by 3 pp for arousal and 2 pp for valence at the RECOLA database. Similarly, at the CreativeIT database, improvements are 2 pp for valence. On the other hand, DCNN has only 0.2 pp better performance for arousal at the CreativeIT database. This consistent improvement indicates that learning relationships at different temporal resolutions improve ABS detection.

KFDCNN has superior performance on the RECOLA than CreativeIT, by approximately 10 pp for arousal and 3 pp for valence. The change in the performance could be because RECOLA is not an acted dataset, and as a result, it includes more

Table 2.7: F-score and Recall performance results of the baseline (FFN), convolutional neural network (CNN), dilated convolutional neural network (DCNN), and the proposed KFDCNN framework for the affective burst detection over the RECOLA and CreativeIT datasets on Arousal (A) and Valence (V) contours

Model	**RECOLA**				**CreativeIT**			
	UAF1		UAR		UAF1		UAR	
	A	V	A	V	A	V	A	V
SVM	51.3	52.3	52.0	54.6	45.0	50.9	53.6	57.7
FFN	56.2	53.7	56.8	54.4	48.1	51.0	55.8	56.1
CNN	59.0	56.3	60.1	59.0	53.8	56.4	58.5	59.6
DCNN	64.3	59.4	64.8	60.3	**57.2**	56.8	**60.3**	58.0
KFDCNN	**67.0**	**61.4**	**68.5**	**62.2**	57.0	**58.5**	59.4	**60.0**

spontaneous changes and exhibits higher mean absolute delta, $|d|$, values within ABSs compared to the CreativeIT.

2.3 Conclusions on Affective Computing

In Chapter 2, we formulate a CER and Affective Burst Detection problem under affective computing.

The proposed CER-MTL framework combines the early fusion of speech and body motion modalities, MTL, and correlation loss. MTL both increases the robustness of the training and performance by forcing to learn shared representations. Moreover, early fusion and MTL bring significant savings in the complexity of the architecture that prevents the overfitting with the limited affective labeled data.

We conducted experiments on CreativeIT and RECOLA databases to capture the effect of window size on the CER, MTL and CCC loss. We found that increasing window size significantly improves the prediction of arousal. The results depict that CER-MTL outperforms CER-STL, especially bringing a 6% CCC improvement in the prediction of valence. In addition, experiments showed that CER-MTL achieves at least 7% and 13% higher CCC values on respectively CreativeIT and RECOLA datasets when the models are trained with the CCC loss rather than the MSE loss. These performance improvements suggest the effectiveness of the proposed system.

With the strong performance of CER-MTL, we showed that affective states can be extracted from videos using uni-and multi-modal architectures. The structure of CER-MTL enables extracting affective-state-related embeddings from hidden states of GRU modules. These embeddings could be used as a latent-representations of audio-visual information that is correlated with arousal, valence, and dominance.

While CER is important and attracted researchers for a long time, Chapter 2 defines affective burst detection as another crucial affective computing problem that can capture affective fluctuations better in the inter-and intra-emotion domain. We address the affective burst detection as an imbalanced binary problem of segmenting affective contour into burst and idle regions.

First, we label the affective contour by first detecting ABPs over the derivatives of the affective attributes. Later, the annotations are generated by extending the ABPs into segment vector ABSs. The proposed KFDCNN architecture is trained with the

generated annotation targets. The conducted experiments show that the KFDCNN outperforms the baseline architecture for F1-score and Recall on both RECOLA and CreativeIT datasets. Moreover, we depicted the importance of dilation and kernel fusion by comparing KFDCNN with CNN and DCNN. It is seen that a larger receptive field size due to dilation brings at least 3 pp improvements, and kernel fusion brings at least 2 pp improvements. Considering these observations, we suggest using KFDCNN for the affective burst detection.

With KFDCNN we showed that affective burst can be detected from audio information efficiently, and can be used along with CER modules to capture salient human behaviors from salient regions in the affective contour.

Chapter 3

AFFECTIVE VIDEO SUMMARIZATION

The internet infrastructure and services have recently matured to the point that any user can record and share unedited videos, resulting in over 500 hours of footage being uploaded per minute to video streaming platforms like YouTube and Twitch. The vast majority of these uploads tend to be user-generated human-centric videos focusing on human activities, interactions, and their emotional and physical attributes [Vicol et al., 2018]. Video summarization, which preserves human-centric representations of emotional and visual elements, looks to be a key challenge for such a massive collection of unedited human-centric videos [Apostolidis et al., 2021, Sun et al., 2016, Zlatintsi et al., 2017].

Summarization of human-centric videos should address the problem of selecting salient video segments or key frames and constructing an informative, interesting, and short summary of the human-centric actions in the video. Emotional changes are frequently associated with notable human-centric actions in the videos. The continuous emotion recognition (CER) problem has been extensively studied in recent literature for tracking the emotional changes as the prediction of affect attributes from audio and visual modalities [Gunes and Pantic, 2010, Trigeorgis et al., 2016, Tzirakis et al., 2021, Köprü and Erzin, 2020]. In affective and emotional settings, high activation and extreme, low or high, valence attributes are expected to appear with informative and interesting human-centric actions. This research aims to look into the relationship between affective information and human-centric actions to improve the human-centric video summarization task.

The coupling of affective computing and video summarization has received little attention in the literature. Two related tracks of study for this coupling appear as: (i) studies modeling the human actions in videos [Bhattacharya et al., 2021, Sun

et al., 2016]; (ii) studies modeling the emotional responses of the viewers [Joho et al., 2011, Money and Agius, 2009].

In recent work for the first track, Bhattacharya et al. use human-centric modalities such as body poses and faces, and model the inter- and intra-relations between humans with graphs on the human-centric videos [Bhattacharya et al., 2021]. In [Sun et al., 2016], Sun et al. address the summarization of the unconstrained videos by first finding salient people in the video, and then creating montages capturing salient actions of these people. Although both studies utilize human actions, which are correlated with the emotional state of humans to an extent, they do not focus on the emotional content or saliency in the emotional domain.

In the second track, video summarization selects key frames based on the physiological responses of the viewers that are expected to be highly correlated with the viewers' emotional states [Joho et al., 2011, Money and Agius, 2009]. In [Joho et al., 2011], the viewers' facial activity is tracked, and then the frames are ranked to extract personal highlights from the videos by using heuristics. Whereas in the other study, the viewers' physiological responses, such as heart rate and electrodermal response, are tracked to analyze sub-segments of the watched videos [Money and Agius, 2009]. These studies leverage physiological information that is highly related to humans' emotional states to generate personal highlights/summaries. On the other hand, these approaches require at least one subject as a video viewer and do not provide feasible and time efficient solutions for automatic video summarization.

In this chapter, we study the use of affective information extracted from visual data to summarize human-centric videos, which we call affective video summarization. We model affective states from the video itself, unlike studies that evaluate viewers' perspectives [Joho et al., 2011, Money and Agius, 2009]. Our model also targets summarization of human-centric videos, but not specifically over visual human-centric modalities as in [Bhattacharya et al., 2021], but using the affective information extracted from the visual data.

For the affective video summarization, we adopt a two-step approach. First, a convolutional recurrent neural network based affective analysis model has been

constructed with the *GoogLeNet* driven with visual features [Szegedy et al., 2015]. Then, we explore affective feature fusion over the affect attributes and affect embeddings together with the temporal and spatial attention mechanisms to define the proposed fully convolutional network based affective video summarization models. To the best our knowledge, this is the first work in literature that combines affective analysis with video summarization for the affective summarization of human-centric videos.

The visual affective analysis model has been trained over the REmote COLlaborative and Affective (RECOLA) dataset [Ringeval et al., 2013]. To show the benefit of our proposed affective video summarization model, we evaluated it on the TvSum [Song et al., 2015], and COGNIMUSE [Zlatintsi et al., 2017] datasets using F-score, face recall, cumulative Kendall's rank correlation coefficient, and Spearman's rank correlation metrics in comparison with four video summarization baselines from the recent literature. The proposed temporal attention-based affective video summarization model is observed to sustain higher performance than the baselines, with consistent improvements for all performance metrics.

To summarize, the main contributions of this study are as follows:

- We formulate a novel end-to-end learning problem for affective-information enriched video summarization targeting human-centric videos.

- A CER model extracts affective information in terms of emotional attributes and learned embeddings from the visual inputs.

- We investigate the use of attention mechanisms in video summarization for human-centric videos and develop temporal and spatial attention-based frameworks.

- We conduct affective video summarization evaluations on the state-of-the-art using both standard and self-defined evaluation metrics.

The rest of Chapter 3 is organized as follows. Chapter 3.1 reviews related work, and Chapter 3.2 describes the main building blocks of the proposed framework.

Chapter 3.3 presents the experiments conducted together with the performance evaluations. Finally, the discussion on affective video summarization is presented in Chapter 3.4.

3.1 Related Work

In Chapter 3, we are investigating the video summarization problem for human-centric videos with the attention to emotional changes in the video content. This brings a novel perspective to the affective video summarization task. In the recent literature, several lines of works relate to this task.

While video summarization is often defined as a supervised sequence to sequence mapping problem over encoder-decoder architectures [Rochan et al., 2018, Ji et al., 2020, Yuan et al., 2019, Li et al., 2021], it has been also formulated under unsupervised settings to maximize a reward function for diversity and representativeness of summaries [Zhou et al., 2018] or to minimize distribution differences between videos and their summarizations [Mahasseni et al., 2017, Apostolidis et al., 2020]. In the supervised approaches, while LSTM models are used to capture long-range dependencies in [Ji et al., 2020, Yuan et al., 2019, Li et al., 2021], Rochan et al. differ from them by processing all frames of subsampled video through a fully convolutional neural network (SUM-FCN) [Rochan et al., 2018]. On the other hand, Zhou et al. formulate the summarization task as sequential decision-making using an end-to-end reinforcement learning-based framework (DR-DSN) [Zhou et al., 2018]. They utilize a reward function that jointly accounts for the diversity and representativeness of the generated summaries in an unsupervised setting. Mahasseni et al. propose a deep LSTM based summarizer and a discriminator network (SUM-GAN) to minimize the distribution of keyframes and the original video in an unsupervised manner [Mahasseni et al., 2017]. The summarizer selects keyframes and then reconstructs the input video, while the discriminator distinguishes the original video from its reconstruction resulting in an adversarial learning framework. Later, Apostolidis et al. introduce an attention mechanism to SUM-GAN by integrating a deterministic attention auto-encoder (SUM-GAN-AAE) [Apostolidis et al., 2020].

Several recent works in video summarization adapt attention to enhance summarization performance [Fajtl et al., 2018, Jung et al., 2020, Ji et al., 2020, Apostolidis et al., 2020]. Early work by Fajtl et al. employs a sequence-to-sequence transfor-

mation for video summarization using a soft self-attention mechanism by defining a frame-level meta-representation as a weighted combination of features from all time frames [Fajtl et al., 2018]. However, the order of information across frames is lost during this process, which could be essential for video summarization. Jung et al. extend the use of self-attention to capture the positional information by introducing global-and-local relative position embeddings [Jung et al., 2020]. In another study, Ji et al. encode visual features with a Bidirectional LSTM network [Ji et al., 2020]. Encoded vectors combine Bahdanau attention [Bahdanau et al., 2015] and then feed into an LSTM based decoder. Although our formulation of video summarization follows a sequence-to-sequence mapping with attention mechanisms, our study differs from these studies in the exploitation of affective information for the video summarization task.

Affective information processing attempts to define and extract attributes of emotion. *Emotions* are defined as mental states that are triggered by neurophysiological changes and are related to feelings and behavioral responses [Ekman and Davidson, 1994]. Behavioral responses co-articulate with speech, bodily gestures, and facial expressions. Hence human-centric videos are expected to carry emotional cues over the behavioral responses [Zlatintsi et al., 2017]. Emotion recognition from audio-visual signals has been extensively studied in the literature [Gunes and Pantic, 2010, Trigeorgis et al., 2016, Köprü and Erzin, 2020]. Emotions are represented both in discrete and continuous domains. Discrete categorical emotions, such as happiness and sadness, can also be represented in the 3-dimensional continuous affect space of activation, valence, and dominance, which are the indicators of activeness-passiveness, positiveness-negativeness, and dominance-submissiveness, respectively [Schlosberg, 1954].

Prediction of activation, valence and dominance attributes from behavioral responses are categorized as continuous emotion recognition (CER) task [Schmitt et al., 2019, Köprü and Erzin, 2020, Tzirakis et al., 2021]. In the recent literature, deep neural network (DNN) based approaches have been extensively used for the CER task. Schmitt et al. investigate arousal and valence prediction from the

low-level descriptors (LLDs) of speech signal using RNNs with the CCC loss function [Schmitt et al., 2019]. Tzirakis et al. design an end-to-end network utilizing raw video, audio, and text modalities [Tzirakis et al., 2021]. They construct a multimodal network having an attention layer to fuse deep embeddings of the text, audio, and visual modalities and an LSTM layer that predicts the activation and valence attributes. In Chapter 2, we performed a feature-level audio-visual information fusion to train a CRNN model with multi-task learning to predict affect attributes.

Although the intersection of affective computing and video summarization is narrow, several studies focus on emotion recognition while either relating their findings to video summarization [Xu et al., 2018, Tu et al., 2020] or utilizing similar ideas like shot segmentation and keyframe selection for emotion recognition [Wang et al., 2021]. Xu et al. investigate information transfer from the image and textual data of videos for emotion recognition, emotion attribution, and emotion-oriented summarization [Xu et al., 2018]. Using zero-shot learning to estimate categorical emotions, they learn video representations over image transfer-encoding and textual representations. Later, they perform emotion attribution to identify the contribution of each frame to the video's overall emotion. Finally, video summarization is formulated as a selection of keyframes by maximizing an emotion attribute-based score function. In the second related study, Tu et al. train a joint model to capture emotion attribution and recognition using multitask learning [Tu et al., 2020]. Later, similar to the previous study [Xu et al., 2018], the video summarization task is formulated as a post-processing optimization problem and solved using MINMAX dynamic programming. Wang et al. segment the videos into shots using audio-visual information and represent each shot with three keyframes [Wang et al., 2021]. Then, they process these keyframes with an action detector, face detector, and person detector while extracting low-level descriptors and high-level features from audio by OpenSmile tool [Eyben et al., 2010] and a pretrained VGGish network. Concatenation of the outputs from these detectors and extractors are then fed into an LSTM with temporal attention to estimate arousal and valence. Note that both [Xu et al., 2018] and [Tu et al., 2020] formulate the summarization task as an optimization problem,

which is solved in post-processing. Hence their video summarization frameworks are not learning-based, and they do not explore how affective information alters the behavior of the proposed summarization architecture. Furthermore, the proposed solutions are not evaluated on the state-of-the-art video summarization datasets. On the other hand, although [Wang et al., 2021] utilizes shot segmentation and key frame selection, they do not include these processes in training; hence these are not updated regarding any saliency in the affective attribute space.

There are also video summarization studies where affect-related attributes or modalities are utilized to generate human-centric video summaries [Koutras et al., 2015, Sun et al., 2016, Bhattacharya et al., 2021]. Koutras et al. extract audio-visual and affective text features to classify each frame as salient or not using a k-nearest neighbors classifier [Koutras et al., 2015]. Affective text features are generated from their distance to selected seed words with known affective attributes. Sun et al. perform a human pose detection to locate people and their trajectories and detect salient human-centric segments using a random forest classifier from the pose, motion, and self-defined camera-centric features [Sun et al., 2016]. Then they generate salient montages representing the video using the human trajectories and saliency scores. In another study utilizing human poses, Bhattacharya et al. represent humans as graphs whose nodes are either 2d face landmarks or 3d body joints of humans [Bhattacharya et al., 2021]. Then using an auto-encoder type of architecture based on a spatial-temporal graph convolutional network, they estimate the importance scores of frames. Although these studies utilize human-centric cues, our study differs from them as we focus directly on extracting affective attributes and high-level emotion embeddings from visual data to locate salient affective regions with attention mechanisms.

3.2 Affective video Summarization Methodology

In this study, we investigate the use of affective information for enriching video summarization by capturing emotionally salient regions of the human-centric videos. First, we state and formulate the video summarization problem and define the visual feature extraction for both the CER and video summarization tasks. Then, an end-to-end framework for the CER is presented. Emotional attributes and high-level embeddings from the CER framework are later used as affective representations by the video summarization framework. Finally, we introduce the proposed affective video summarization architectures by first defining a video summarization baseline and then enriching this baseline with the fusion of affective information.

3.2.1 Problem Statement

Video summarization is widely formulated as either a binary classification or a frame-level regression task. In the binary classification task, summarization outputs are either key-frames [Rochan et al., 2018, Zhang et al., 2016] or key-shots [Song et al., 2015, Zhang et al., 2016] from the video. On the other hand, frame-level importance scores are extracted in the regression task [Ji et al., 2020, Zhang et al., 2016].

In this study, we formulate the video summarization as a binary classification task where the positive labels correspond to the selected key-frames. The summarization network receives a visual feature matrix, $\mathbf{v} \in \mathbb{R}^{N \times D}$, and emits an output matrix, $\mathbf{s} \in \mathbb{R}^{N \times C}$, where N is the number of frames in the video, D is the dimensionality of the frame-level visual feature, and C is the number of classes. We take $C = 2$ representing the positive and negative classes for the key-frame selection and these two nodes output class probability values at the output of the network. Then the positive class for key-frame selection or the negative class for frame skip are set by picking the node with higher probability. Eventually, the video summary is constructed from the key-frames that are labeled as positive.

In this study, we extract the visual information using the *GoogLeNet* [Szegedy et al., 2015]. Output of the *pool5* layer of the pre-trained *GoogLeNet* is used as the

visual feature and represented as $\mathbf{e}_i \in \mathbb{R}^D$ at frame i with dimension $D = 1024$.

3.2.2 CER Network (CER-NET)

The continuous emotion recognition problem is set as the continuous regression of an emotional attribute from the temporal visual features. For this purpose, we construct a CER network (CER-NET), which consists of two back-to-back convolutional layers, a max pool layer in the temporal domain, a Gated Recurrent Unit (GRU) layer, and a fully connected layer as shown in Figure 3.1. We use the CER-NET to train two separate networks to estimate the activation and valence (AV) attributes separately. A temporal visual feature matrix is defined to be the input of the CER-NET. For this purpose, visual features around the i-th frame are cascaded to define the temporal visual feature as

$$\mathbf{f}_i^{\text{CER}} = [\mathbf{e}_{i-\Delta+1}, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_i, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+\Delta}] \in \mathbb{R}^{D \times T} \tag{3.1}$$

at frame i, where T is the temporal window size and it is set as $T = 2\Delta = 20$ frames.

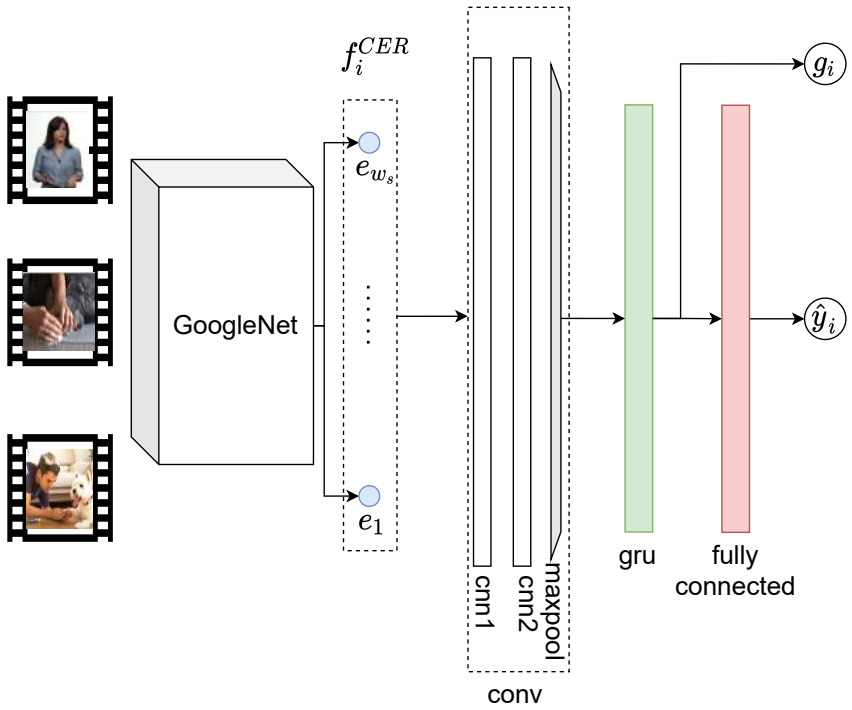


Figure 3.1: CER-NET: The continuous emotion recognition network

We refer the group of back-to-back two convolutional and a max-pool layers as the *conv* layer for the sake of simplicity. In CER-NET, the *conv* layer models the spatial relations and provides a compact representation of the temporal window. In this manner, both temporally related features are highlighted and dimensionality of the representation is reduced. Dimensionality reduction is the key to complexity reduction and it also prevents overfitting. The compact representation from the *conv* layer is fed into GRU to model long-term temporal relations.

At the inference phase, CER-NET receives $\mathbf{f}_i^{\text{CER}}$ and provides the GRU layer outputs as $\mathbf{g}_i \in \mathbb{R}^G$ and the fully connected layer outputs as $\hat{y}_i \in \mathbb{R}$. We define the GRU layer output, $\mathbf{g}_i$, as an affective embedding, which can carry affective information to the video summarization model. On the other hand, the fully connected layer output, $\hat{y}_i$, represents the estimated emotional attribute (A or V) at frame i and delivers another source of affective information.

3.2.3 Affective Video Summarization

The proposed affective video summarization (AVSUM) architecture is based on a fully convolutional neural network (FCN) for semantic segmentation [Long et al., 2014], which is adapted by [Rochan et al., 2018] into video summarization (SUM-FCN). In the following, we briefly describe SUM-FCN and then present two fusion architectures to combine affective information with video summarization, which are later extended by temporal and spatial attention mechanisms.

SUM-FCN

SUM-FCN adopts an encoder-decoder architecture where the encoder is based on fully convolutional layers and the decoder is based on deconvolutions. Figure 3.2 includes the architecture of the SUM-FCN compromising encoder-bottleneck-decoder layers driven by the visual input $\mathbf{v}$. The encoder compresses the temporal information while increasing spatially, and the bottleneck passes it to the decoder to reconstruct necessary information from this compact representation. SUM-FCN receives the visual features for the whole video at once and provides the summa-

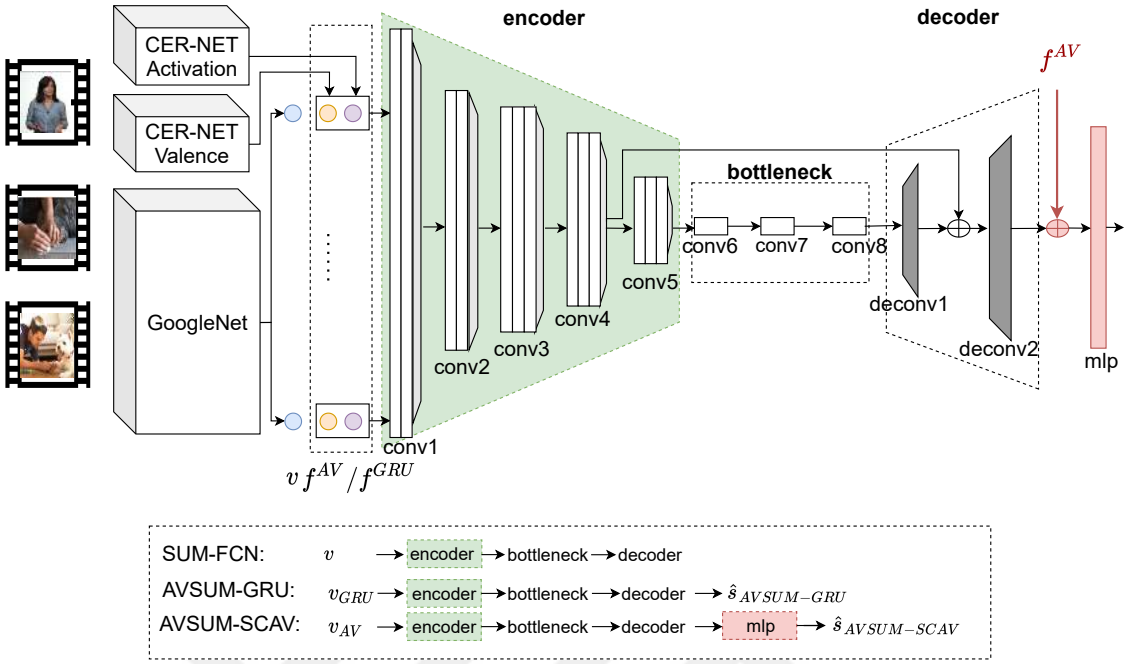


Figure 3.2: An overview of the proposed AVSUM-GRU and AVSUM-SCAV architectures. AVSUM-GRU combines high-level affective information f_{GRU} from the CER-NET estimators with the GoogleNet drived visual features. AVSUM-SCAV defines a skip-connection for the emotional attributes f_{AV} as well as combining them with the GoogleNet drived visual features for the video summarization.

rization outputs at once. Video frame rate is typically down-sampled before the summarization. Let $\mathbf{v}$ be the input of SUM-FCN, then $\mathbf{v} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]' \in \mathbb{R}^{N \times D}$ where N is number of frames in the down-sampled stream. Given $\mathbf{v}$, SUM-FCN emits $\mathbf{s} \in \mathbb{R}^{N \times C}$, where C is the dimension of the summarization annotations and set as $C = 2$ as defined in section 3.2.1.

Affective Feature Fusion for AVSUM

Affective information is extracted from the two CER-NET models, which are trained to estimate the activation and valence (AV) attributes separately. Two types of affective information cues are extracted from the CER-NET models: (i) the estimated AV attributes ($\mathbf{f}^{\text{AV}}$), and (ii) the learned high-level CER-NET representations based on the GRU embeddings ($\mathbf{g}^{\text{AV}}$). The estimated AV attribute vector $\mathbf{f}_j^{\text{AV}}$ is con-

structed as a column vector from the estimated activation and valence attributes as $\mathbf{f}_j^{AV} = [\hat{y}_j^A, \hat{y}_j^V]' \in \mathbb{R}^2$ at frame j in the down-sampled stream. Similarly, $\mathbf{g}_j^{\text{AV}}$ is constructed by concatenating the outputs of the GRU layers from the two CER-NET models as $\mathbf{g}_j^{\text{AV}} = [\mathbf{g}_j^{\text{A}\prime}, \mathbf{g}_j^{\text{V}\prime}]' \in \mathbb{R}^{2G}$. Then, the emotional attribute $\mathbf{f}^{\text{AV}}$ and the affect embedding $\mathbf{f}^{\text{GRU}}$ representations of the video are defined as

$$\mathbf{f}^{\text{AV}} = [\mathbf{f}_1^{\text{AV}}, \ldots, \mathbf{f}_N^{\text{AV}}]' \in \mathbb{R}^{N\times 2} \tag{3.2}$$

$$\mathbf{f}^{\text{GRU}} = [\mathbf{g}_1^{\text{AV}}, \ldots, \mathbf{g}_N^{\text{AV}}]' \in \mathbb{R}^{N\times 2G}. \tag{3.3}$$

The first proposed AVSUM architecture, referred as AVSUM-GRU, combines the affect embedding $\mathbf{f}^{\text{GRU}}$ with the visual input $\mathbf{v}$. Figure 3.2 presents the AVSUM-GRU architecture receiving the $\mathbf{v}_{\text{GRU}}$ input as

$$\mathbf{v}_{\text{GRU}} = \mathbf{v} \oplus \mathbf{f}^{\text{GRU}} \in \mathbb{R}^{N\times(D+2G)}, \tag{3.4}$$

where $\oplus$ operator is representing the feature vector combining over the whole video.

Alternatively, we define the AVSUM-SCAV architecture, as illustrated in Figure 3.2, by combining the emotional attribute $\mathbf{f}^{\text{AV}}$ with the visual input $\mathbf{v}$ to receive $\mathbf{v}_{AV}$ as

$$\mathbf{v}_{\text{AV}} = \mathbf{v} \oplus \mathbf{f}^{\text{AV}} \in \mathbb{R}^{N\times(D+2)} \tag{3.5}$$

and incorporating a long skip-connection by concatenating the emotional attribute $\mathbf{f}^{\text{AV}}$ to the final layer of the summarization network. AVSUM-SCAV inserts a skip-connection to the output of *deconv2* layer and has a final fully connected *mlp* layer to reduce the dimension from $C + 2$ to C.

Temporal Attention for AVSUM

We adapt multi-headed attention (MHA) mechanism into the AVSUM architecture to efficiently model temporal dependencies across the video frames and referred as TA-AVSUM. Similar to AVSUM-GRU, TA-AVSUM receives affective information enriched features $\mathbf{v}_{\text{GRU}}$ as the input, and in addition to AVSUM-GRU it employs attention to the temporally compressed sequence. Figure 3.3 depicts the MHA

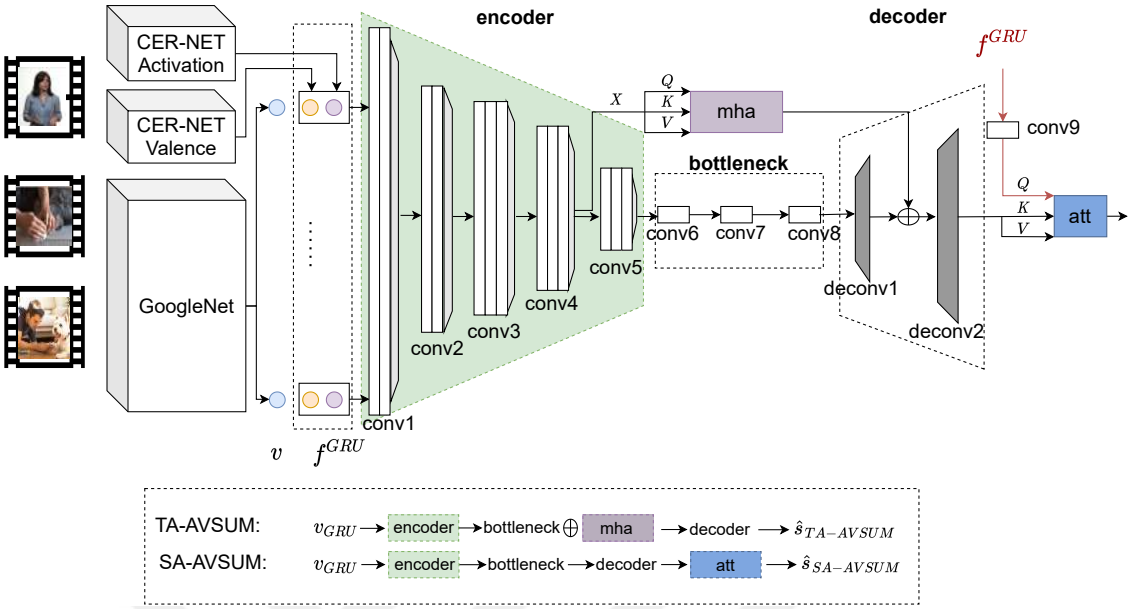


Figure 3.3: An overview of the proposed TA-AVSUM and SA-AVSUM architectures. TA-AVSUM applies multi-head-self attention to the output of *conv4* layer (X), on top of AVSUM architecture. On the other hand, SA-AVSUM modifies the long skip connection of affective features by applying a spatial attention.

based temporal attention structure, where MHA is placed to the output of *conv4* layer which emits $\mathbf{X} \in \mathbb{R}^{M \times S}$. Here, M and S are respectively the temporal and spatial dimensions of $\mathbf{X}$.

MHA receives three inputs as Query ($\mathbf{Q}$), Key ($\mathbf{K}$) and Value ($\mathbf{V}$), then outputs a weighted summation of the rows of $\mathbf{V}$. The weights are calculated from the similarity between the $\mathbf{Q}$ and $\mathbf{K}$. In this context, the MHA is defined as

$$\text{head}_h = \text{softmax}(\frac{\mathbf{Q}\mathbf{W}_h^Q(\mathbf{K}\mathbf{W}_h^K)^T}{\sqrt{M}})\mathbf{V}\mathbf{W}_h^V \tag{3.6}$$

$$\text{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_H)\mathbf{W}^O, \tag{3.7}$$

where $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, $\mathbf{W}_h^V$ and $\mathbf{W}^O$ are the learned linear projections and H is the number of heads. We employ multi-headed self-attention, by setting $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ to $\mathbf{X}$. Hence, the learned projections matrices are formed as $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, $\mathbf{W}_h^Q \in \mathbb{R}^{S \times \frac{S}{H}}$, and $\mathbf{W}^O \in \mathbb{R}^{M \times M}$.

Spatial Attention for AVSUM

Motivated by the Squeeze and Excitation Networks [Hu et al., 2018], which attend to different channels of an image, we propose the fourth AVSUM network with spatial domain attention and refer it as SA-AVSUM. Figure 3.3 depicts the SA-AVSUM structure where we employ attention to the output of the *deconv2* layer. Unlike TA-AVSUM, we adopt single-headed attention, which is formulated in (3.6). Then, $\mathbf{K}$ and $\mathbf{V}$ are set to transpose of $\hat{\mathbf{s}}_{\text{AVSUM}}$ and $\mathbf{Q}$ is set from the affective embedding $\mathbf{f}^{\text{GRU}}$.

3.2.4 Model Training

Training of the CER-NET and video summarization models are executed in two phases. First the CER-NET models are trained, later they are fixed and integrated for the affective video summarization to train the AVSUM-GRU, AVSUM-SCAV, TA-AVSUM, and SA-AVSUM models.

CER-NET

The CER-NET models are trained separately to estimate the activation and valence attributes using the concordance correlation coefficient (CCC) based loss function. The loss function of the CER-NET is defined as the negated CCC value,

$$L_{\text{CER}} = -\frac{2\sigma_{y\hat{y}}^2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{3.8}$$

where y is the ground truth and $\hat{y}$ is the estimated attribute.

Video Summarization Networks

The key frame selection problem has an imbalanced nature, since only a small number of frames are selected for the summary [Rochan et al., 2018]. In order to overcome the imbalance problem, a weighted binary cross entropy loss is defined as

$$L_{\text{SUM}} = -\frac{1}{N}\sum_{j=1}^{N} w_{z_j}(z_j \log \hat{z}_j + (1 - z_j)\log(1 - \hat{z}_j)), \tag{3.9}$$

where $z_j \in 0, 1$ is the binary ground truth, $\hat{z}$ is the predicted score and w_{z_j} is the weight of the j^{th} frame. The weights for the binary target are defined as

$$w_0 = \frac{1}{N} \sum_{j=1}^{N} z_j \quad \text{and} \quad w_1 = 1 - w_0. \tag{3.10}$$

3.3 Experiments

Experimental evaluations of the proposed models and comparisons with the state-of-the-art are performed using two datasets. In this section, we first introduce the datasets and evaluation metrics, then explain implementation details. Finally, experimental results are presented and discussed.

3.3.1 Datasets

We train and evaluate the CER-NET on the RECOLA dataset, which is a popular multi-modal dataset for emotion recognition [Ringeval et al., 2013]. The RECOLA dataset is composed of multi-modal recordings of dyadic conversations from 27 French speakers. From these 27 recordings, 18 of them are annotated, and the rest of the records are used for testing. The annotations are at the rate of 40 msec and from 6 different annotators. In total, we use 90 minutes of recordings from the RECOLA dataset in this study.

Experimental evaluations on the proposed AVSUM architectures are executed on the frequently used TvSum [Song et al., 2015], and COGNIMUSE [Zlatintsi et al., 2017] datasets. Note that there is no available video summarization dataset containing only human-centric videos in the literature. The TvSum dataset contains 50 user generated videos from 10 different categories, such as vehicle tire changing, sandwich making, grooming an animal, etc. The demographics of the dataset in terms of the face including frames in summary and the full video clip, are presented in Figure 3.4. We categorize videos into human-centric and rest regarding the number of faces in summary, where we set the threshold to 80 frames labeling 15 videos as human-centric. The frame-level importance scores are provided as ground truths for each video from 20 raters. We follow the approach in [Zhang et al., 2016, Rochan et al., 2018] to convert the frame-level importance scores into the key shot-based summaries.

The COGNIMUSE dataset contains half-hour continuous segments from six Hollywood movies, and is annotated with sensory and semantic saliency, audio and vi-

sual events, cross-media relations as well as emotion [Zlatintsi et al., 2017]. Emotions are annotated at frame level for arousal and valence as continuous values ranging between -1 and 1. Saliency annotations are generated from audio, video, semantics, audio-visual, and audio-visual semantics as binary labels. Saliency ground truths are provided as a single stream generated from agreed regions of 3 annotators. In this study, we utilize ground truths from the visual only medium.

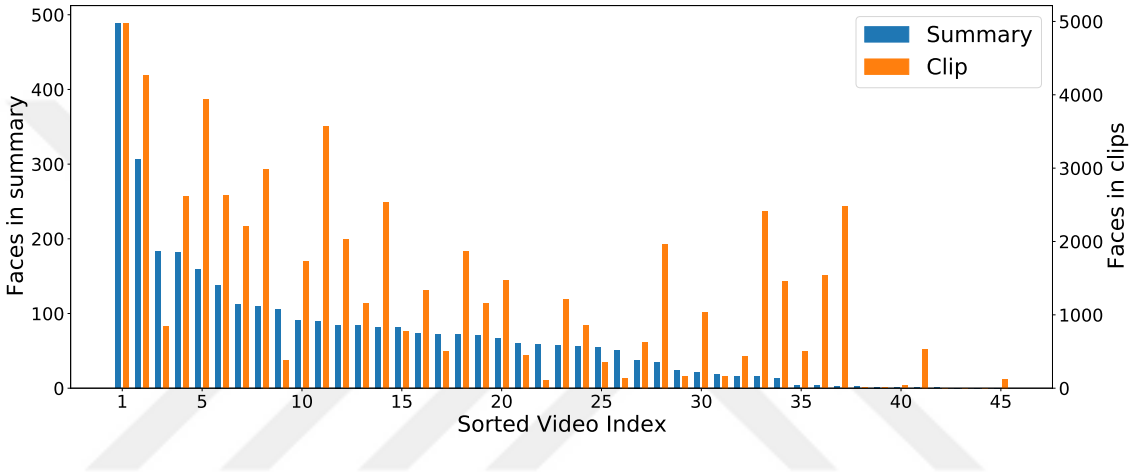


Figure 3.4: Number of face including frames in the summary and total video for TvSum dataset where the videos are sorted regarding the number of faces in the summary.

3.3.2 Evaluation Metrics

Emotional attribute estimation with the CER-NET is evaluated using the CCC metric following [Köprü and Erzin, 2020], which is the negative of the L_{CER} loss in (3.8).

For the video summarization task, following [Rochan et al., 2018, Zhang et al., 2016, Ji et al., 2020], F-score is used as an evaluation metric. For the k-th video, let the ground truth binary summarization vector be $\mathbf{s}^k$ and estimated binary summarization vector to be $\hat{\mathbf{s}}^k$ where $\mathbf{s}^k, \hat{\mathbf{s}}^k \in \mathbf{R}^N$. Then precision P^k and recall R^k for the k-th video are calculated as

$$P^k = \frac{\mathbf{s}^k \cdot \hat{\mathbf{s}}^k}{\sum_j^N \hat{s}_j^k} \quad \text{and} \quad R^k = \frac{\mathbf{s}^k \cdot \hat{\mathbf{s}}^k}{\sum_j^N s_j^k}, \tag{3.11}$$

where j runs over the frames. Video level F-score is defined as the harmonic mean of the precision and recall,

$$F1^k = \frac{2P^k R^k}{P^k + R^k}. \tag{3.12}$$

Then, the final F-score metric $F1$ is defined as the unweighted average of the video level F-score values as

$$F1 = \frac{1}{K} \sum_{k=1}^{K} F1^k, \tag{3.13}$$

where K is the number of videos in the dataset.

Since the affective information is learned from a dataset of human-centric videos, capability of capturing affective frames of the video during the summarization is expected to be better with human-centric videos. By highlighting this fact, we define two new metrics for the evaluation of the affective video summarization. The first metric computes normalized F1 score differences with the baseline over the videos with the highest number of face appearances. To define this metric, let us first define the normalized F1 score difference for the k-th video with respect to the SUM-FCN baseline model as

$$\Delta F1^k = \frac{F1^k - F1^k_{SUM-FCN}}{F1^k_{SUM-FCN}}, \tag{3.14}$$

where $F1^k$ refers to the F1 score of the model in evaluation. Also assume that all the videos in the dataset are sorted with the highest number of face appearances in descending order and are indexed with k_l for $l = 1, \ldots, K$. Then, the cumulative F1 score difference metric for the *Top-L* human-centric videos is defined as

$$\Delta F1_L = \sum_{l=1}^{L} \Delta F1^{k_l}. \tag{3.15}$$

Associated with the $\Delta F1_L$, we also compute the F1 score of the *Top-L* human-centric videos and refer it as $F1_L$.

Human-centric nature of the videos can be associated with the face appearances in the video frames. Motivated with this fact, we set a second metric for evaluation of the affective video summarization as the recall rate of face appearing frames in

the extracted summary. Let $\mathbf{d}_F^k$ be the binary vector representing whether a frame includes a face appearance or not for the k-th video. Binary face appearance vectors are extracted by the histogram of oriented gradients (HOG) based face detector [King, 2009]. Then, the recall rate of face appearing frames, let's refer it as face recall, R is defined as

$$R = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\mathbf{s}}^k \cdot (\mathbf{d}_F^k \odot \mathbf{s}^k)}{\sum_j^N s_j^k}, \tag{3.16}$$

where $(\mathbf{d}_F^k \odot \mathbf{s}^k)$ is the Hadamard product of $\mathbf{d}_F^k$ and $\mathbf{s}^k$.

Following [Otani et al., 2019], to evaluate the predicted importance scores $(\hat{z})$, we utilize Kendall's rank correlation τ and Spearman's rank correlation coefficient (ρ). Let ζ^k be the reference importance scores, and z^k be the predicted scores for the k-th video, then $(\zeta_i^k, \hat{z}_i^k)$ and $(\zeta_j^k, \hat{z}_j^k)$ are concordant pairs if either both $\zeta_i^k > \zeta_j^k$ and $\hat{z}_i^k > \hat{z}_j^k$ holds or both $\zeta_i^k < \zeta_j^k$ and $\hat{z}_i^k < \hat{z}_j^k$, else $(\zeta_i^k, \hat{z}_i^k)$ and $(\zeta_j^k, \hat{z}_j^k)$ are considered as discordant pairs. Based on the concordant and discordant pairs, Kendall's τ is defined as

$$\tau = \frac{1}{N}\sum_{k=1}^{K}\frac{N_c^k - N_d^k}{\sqrt{(N_c^k + N_d^k + T^k)(N_c^k + N_d^k + U^k)}}, \tag{3.17}$$

where N_c^k, N_d^k, T^k and U^k are the number of concordant pairs, discordant pairs, ties in ζ^k, and ties in $\hat{z}^k$ for the k-th video.

Spearman's rank correlation coefficient is defined as

$$\rho = \frac{1}{N}\sum_{k=1}^{K}\frac{cov(R_{\zeta^k}, R_{\hat{z}^k})}{\sigma_{R_{\zeta^k}}\sigma_{R_{\hat{z}^k}}} \tag{3.18}$$

where R_{ζ^k} and $R_{\hat{z}^k}$ are the ranks, $cov(R_{\zeta^k}, R_{\hat{z}^k})$ is the covariance between R_{ζ^k} and $R_{\hat{z}^k}$, and $\sigma_{R_{\zeta^k}}$ and $\sigma_{R_{\hat{z}^k}}$ are the standard deviations of R_{ζ^k} and $R_{\hat{z}^k}$ respectively for the k-th video.

Similar to $F1_L$, we compute the rank correlations of *Top-L* human centric videos and refer them as τ_L and ρ_L.

We also study statistical differences of a given affective feature dimension, f, across face appearing and not-appearing frames. For this purpose, Kullback-Leibler

(KL) divergence of f in these two classes is defined as

$$D(P_f||Q_f) = \sum_{x \in \mathcal{X}} P_f(x) \log(\frac{P_f(x)}{Q_f(x)}), \tag{3.19}$$

where P_f and Q_f are respectively probability distributions of the affective feature dimension f across face appearing and face not-appearing frames. Note that the affective feature dimension f is driven from the affective feature set as $f \in \{f^A, f^V, g_1^{AV}, g_2^{AV}, \ldots, g_{2G}^{AV}\}$.

3.3.3 Implementation Details

CER-NET

The window size T is selected as 20 frames. The *cnn1* and *cnn2* layers have 10 filters, max-pooling layer reduces the temporal dimension from 20 to 5, *gru* layer has 10 cells and *FCN* layer has 1 node. Hence, the dimensionality of $\mathbf{f}^{\text{AV}}$ is 2 and $\mathbf{g}_{\text{AV}}$ is 20 ($G = 10$).

Annotated recordings of the RECOLA dataset are used during the training of the CER-NET. RECOLA recordings are divided as 10% for the test, 10% for the validation, and 80% for the training. We applied Adam optimizer with a learning rate of 10^{-4} and with a batch size of 256.

Video Summarization Networks

Following [Rochan et al., 2018, Zhang et al., 2016], TvSum videos are downsampled to 2 fps, and frames are fed into GoogLeNet. For the AVSUM-SCAV, the input feature dimension is set as 1026. The input dimension of the AVSUM-GRU, TA-AVSUM, and SA-AVSUM models became 1044 with the GRU embeddings. *conv1* layer has For the TA-AVSUM, we set the number of heads $H = 4$.

Two convolutional layers in *conv1* and *conv2* have 1044 filters and a kernel size of 3. The three convolution layers in *conv3*, *conv4*, and *conv5* have a filter size of 3 and a number of filters of 1044, 2088, and 2088 respectively. The *conv6*, *conv7*, *conv8*, and *conv9* have a filter size of 1 and the number of filters of 4196, 4196, 2,

and 2 respectively. All the max-pool layers have a kernel size of 2 and stride length of 2. Finally, the *deconv1* block upsamples the input by 4, while the *deconv2* block upsamples the input by 16.

We adopted the leave one-group out cross-validation technique to compare the performances. Nine videos are selected from the TvSUM dataset for each fold, and the rest are used for training. This procedure results in five folds where non-of-these-test sets intersect; hence each video is included once at the test set. One video is selected for the cross-validation evaluation on the COGNIMUSE, and the rest is used for training, resulting in six-folds. Results of the state-of-art models are also generated following the described cross-validation procedures.

We mimic fixed size cropping in semantic segmentation by uniformly sampling the video frames and using $N = 320$ for TvSUM [Rochan et al., 2018] and $N = 1280$ for COGNIMUSE dataset. The training is held for 50 epochs with a batch size of 5 videos. During the training phase, Adam optimizer with the learning rate of 10^{-3} is used. For each training fold, a model achieving the highest F-score ($F1$) and a model achieving the highest face recall (R) are selected for the performance evaluations.

3.3.4 CER-NET Performance

Table 3.1 presents the CCC performance of the CER-NET emotion recognition model driven by the visual modality on the RECOLA dataset. The proposed architecture performs better at estimating the activation than the valence. The end-to-end CER-NET architecture outperforms CER-MTL Facial model [Köprü and Erzin, 2020], AVEC-2016 baseline [Valstar et al., 2016] and End-to-end visual network [Trigeorgis et al., 2016] in estimating activation by achieving CCC of 0.40. The end-to-end visual networks [Trigeorgis et al., 2016, Tzirakis et al., 2018] and AVEC-2016 baseline [Valstar et al., 2016] perform strongly for the valence estimation and fairly close to the CER-NET for the activation estimation. Different than CER-NET, [Köprü and Erzin, 2020] and [Valstar et al., 2016] receives visual information as facial activation units and optical flow vectors. Due to the input dimension, CER-NET has more trainable coefficients leading to a more complex structure than the

CER-MTL Facial. Also note that End2You [Tzirakis et al., 2018] and AVEC-2016 baseline [Valstar et al., 2016] results are reported after a post-processing of the affect attributes with median filtering, centering, scaling, and time-shifting, which are not used in the CER-NET evaluations.

Table 3.1: CCC performance of the CER-NET emotion recognition model on the RECOLA dataset

Model	**Activation**	**Valence**
CER-NET	0.40	0.17
AVEC-2016 Video-geometric [Valstar et al., 2016]	0.38	0.61
CER-MTL Facial [Köprü and Erzin, 2020]	0.15	0.06
End2You [Tzirakis et al., 2018]	0.47	0.58
End-to-end Visual Network [Trigeorgis et al., 2016]	0.36	0.48

In comparison with these baseline visual models [Köprü and Erzin, 2020, Trigeorgis et al., 2016, Valstar et al., 2016, Tzirakis et al., 2018], CER-NET performs competitively in representing affective information on the visual channel.

3.3.5 Cumulative AVSUM Performance

This section presents performance evaluations of the proposed affective video summarization models in terms of cumulative F-score, face recall, and rank correlation metrics. Then, the video level performances are investigated to better highlight characteristics of videos that have improved summarization performance with the affective cues.

Table 3.2 presents cumulative F-score ($F1$) and face recall (R) together with the *Top-15* F-score ($F1_{15}$) and face recall (R_{15}) performances for the proposed AVSUM and the baseline models over the TvSUM and COGNIMUSE datasets. In each column, the top-two scoring performances are highlighted in bold. Recall that we

apply two model selection criteria based on F-score and face recall maximization. Each model selection criterion is observed to favor its related performance metric in the evaluations. That is, maximization of $F1$ (Max $F1$) yields higher F-score while maximization of R (Max R) yields higher face recall R.

Table 3.2: Cumulative F-score ($F1$) and face recall (R) together with the *Top-15* F-score ($F1_{15}$) and face recall (R_{15}) performances of the AVSUM and the state-of-art models with the maximization of $F1$ and R model selection criteria over the TvSUM and COGNIMUSE datasets (top two performances for each metric are in bold)

Model	TvSUM								COGNIMUSE			
	Max $F1$				Max R				Max $F1$		Max R	
	$F1$	$F1_{15}$	R	R_{15}	$F1$	$F1_{15}$	R	R_{15}	$F1$	R	$F1$	R
SUM-FCN [Rochan et al., 2018]	57.46	60.00	53.20	**65.52**	54.10	57.40	60.62	60.28	47.55	55.21	46.24	57.73
DR-DSN [Zhou et al., 2018]	56.46	57.54	48.87	60.56	**56.46**	57.54	48.87	60.56	46.19	51.81	46.19	51.81
SUM-GAN [Mahasseni et al., 2017]	56.74	58.14	50.19	61.78	**55.51**	58.06	53.07	68.38	46.09	52.72	45.35	54.59
SUM-GAN-AAE [Apostolidis et al., 2020]	57.37	58.81	45.33	59.87	54.73	57.38	50.74	**69.79**	47.45	53.05	45.32	**64.76**
AVSUM-GRU	**57.50**	**60.25**	**54.04**	59.14	54.02	57.40	60.77	66.20	**47.56**	**55.46**	**46.58**	57.78
AVSUM-SCAV	56.64	**60.60**	52.11	63.80	53.80	57.42	**62.06**	59.64	**47.59**	55.26	46.40	57.00
TA-AVSUM	57.47	59.95	**53.22**	**65.55**	53.79	**59.23**	**65.12**	**70.31**	47.54	**58.16**	**46.41**	**59.00**
SA-AVSUM	55.92	59.98	49.25	62.38	52.76	**59.90**	58.85	62.55	47.01	55.22	46.34	56.8

Observing the cumulative $F1$ performances, the AVSUM-GRU model is competitive for both model selection criteria and observed as the best performing model on the TvSum dataset and the second-best performing model on the COGNIMUSE dataset with the Max $F1$ criterion. On the other hand, the cumulative R score highlights AVSUM-GRU and the temporal attention-based TA-AVSUM models with the Max $F1$ criterion on both datasets. On the other hand, it highlights AVSUM-SCAV and TA-AVSUM models with the Max R criterion on the TvSum dataset. The temporal attention-based TA-AVSUM model especially performs significantly better with the Max R criterion achieving a 65.12% face recall rate. TA-AVSUM's high achieving R score performance consolidated on the COGNIMUSE dataset. TA-AVSUM is observed as the best model on R score with Max $F1$ criterion and the second-best model with Max R criterion on the COGNIMUSE dataset.

Table 3.3 depicts the evaluation of the predicted importance scores with Kendall's

τ and Spearman's ρ for Max $F1$ and Max R. Similar to Table 3.2, the top two performing scores are highlighted on each column. Observing Kendall's τ and Spearman's ρ performances for both of the model selection criteria, TA-AVSUM sustains a place in the top-two performing scores for all columns. This finding aligns with the TA-AVSUM's high performing F-score and face recall scores in Table 3.2. In addition to TA-AVSUM, SA-AVSUM has competitive results with the Max $F1$ criterion, and it outperforms the other architectures with the Max R criterion.

Table 3.3: Cumulative Kendall's τ and Spearman's correlation ρ together with the *Top-15* τ (τ_{15}) and Spearman's correlation (ρ_{15}) performances of the AVSUM and the state-of-art models with the maximization of $F1$ and R model selection criteria over the TvSUM dataset (top two performances for each metric are in bold)

Model	**TvSUM**							
	Max $F1$				**Max** R			
	τ	τ_{15}	ρ	ρ_{15}	τ	τ_{15}	ρ	ρ_{15}
SUM-FCN [Rochan et al., 2018]	**0.012**	0.012	0.016	0.016	0.012	0.013	0.016	0.017
DR-DSN [Zhou et al., 2018]	-0.001	-0.008	-0.001	-0.011	-0.001	-0.008	-0.001	-0.011
SUM-GAN [Mahasseni et al., 2017]	-0.024	0.010	-0.032	0.012	-0.042	0.010	-0.055	0.012
SUM-GAN-AEE [Apostolidis et al., 2020]	0.010	**0.047**	0.013	**0.061**	-0.034	0.001	-0.045	0.001
AVSUM-GRU	0.012	0.013	0.015	0.015	0.010	0.014	0.013	0.018
AVSUM-SCAV	0.011	0.011	0.015	0.014	0.013	0.018	0.17	0.023
TA-AVSUM	**0.013**	**0.014**	**0.017**	**0.018**	**0.015**	**0.019**	**0.020**	**0.024**
SA-AVSUM	0.012	0.014	**0.016**	0.018	**0.016**	**0.031**	**0.022**	**0.041**

Affective information is modeled with the continuous emotion recognition task trained on the RECOLA dataset. Since RECOLA is a human-centric dataset, which includes facial videos, we choose to evaluate the video summarization performance for the *Top-L* human-centric videos, where L is set as 15 with the discussion in Section 3.3.1. Table 3.2 presents the *Top-15* F-score $F1_{15}$ and face recall R_{15} performances on the TvSUM dataset. While attention-based SA-AVSUM and TA-AVSUM models perform best for the $F1_{15}$ score with the Max R criterion, AVSUM-SCAV and AVSUM-GRU models perform best with the Max $F1$ criterion. Overall, the

best performance is 60.60% $F1_{15}$ score with the Max $F1$ criterion for the AVSUM-SCAV model. Observing the *Top-15* face recall R_{15} performances, while TA-AVSUM model is competitive with the baseline SUM-FCN model with the Max $F1$ criterion, it performs significantly superior with the Max R criterion achieving 70.31% face recall R_{15} rate.

Table 3.3 presents Kendall's τ_{15} and Spearman's ρ_{15} performances over the *Top-15* human-centric videos on the TvSUM dataset. Similar to $F1_{15}$ with Max R criterion, attention-based SA-AVSUM and TA-AVSUM outperform the rest for τ_{15} and ρ_{15} with Max R criterion. These results are aligned with their $F1_{15}$ performance and show that attention models are successful in fusing the affective information. On the other hand, SUM-GAN-AEE [Apostolidis et al., 2020] outperforms proposed architectures with the Max-$F1$ criterion, while TA-AVSUM achieves the second-best results. Interestingly, GAN-based models gain significant performance on human-centric videos, which could be due to their ease at generating and discriminating shared human-centric content across the videos.

Table 3.2 highlights two runner up models, AVSUM-GRU and TA-AVSUM, while Table 3.3 consolidates the strong performance of TA-AVSUM. AVSUM-GRU model sustains strong F-score performance with the Max $F1$ criterion, especially for human-centric videos targeted with $F1_{15}$ performance. Alternatively, while temporal attention based TA-AVSUM model performs strongly for the face recall with the Max R criterion and attains 70.31% face recall R_{15} rate, it also sustains a competitive performance for the F-score and face recall rates with the Max $F1$ criterion.

The performance results in Table 3.2 and Table 3.3 show that the TA-AVSUM model is a strong runner up, indicating a strong integration of temporal attention with affective information for affective video summarization.

3.3.6 Explainability

We conduct explainability evaluations to better understand contributions of the affective information and the proposed model architectures for the affective video summarization. First, we present KL divergence analysis for the affective feature

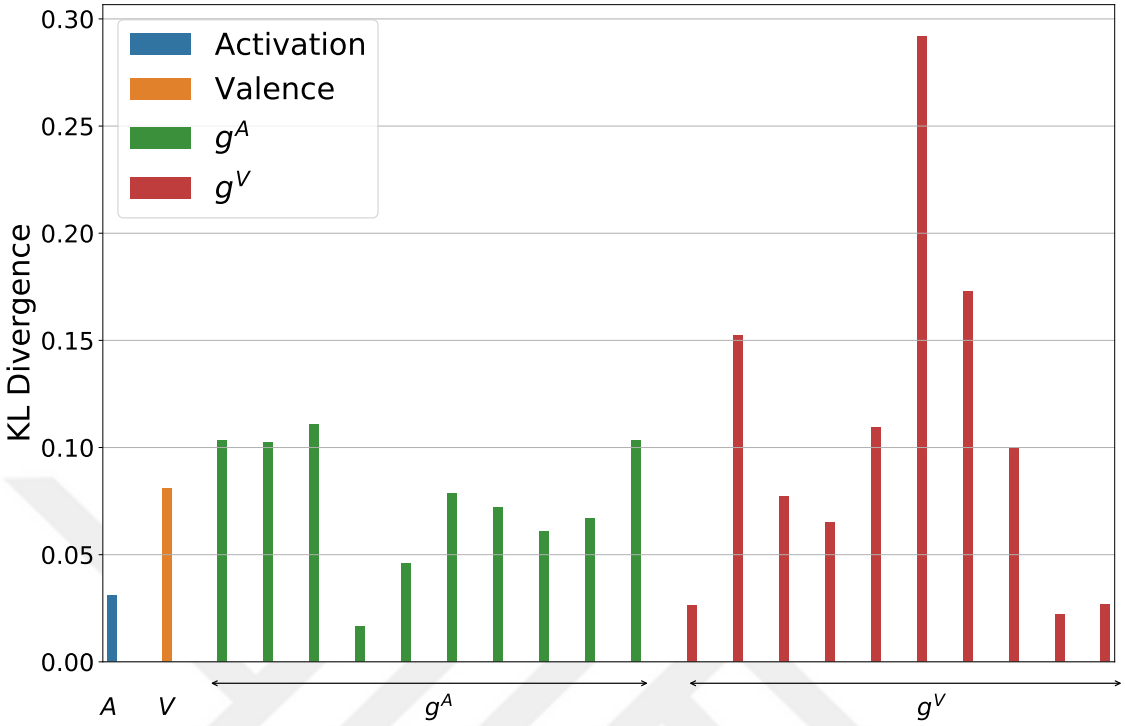


Figure 3.5: KL divergence $D(P_f||Q_f)$ of each affective feature dimension

dimensions. Then, we investigate video level summarization performances of the AVSUM models in terms of the cumulative F-score difference metric $\Delta F1_L$ for the *Top-L* human-centric videos.

Affective Feature Evaluations

Figure 3.5 depicts the KL divergence (KLD) of the affective features across distributions gathered on the frames with faces and without faces. A higher KL divergence indicates bigger discrimination for the distributions of the feature dimension across the with/without face classes. In Figure 3.5, the first two KLD values are for the activation and valence attributes, and the later values are color coded for the dimensions of $\mathbf{g}^A$ and $\mathbf{g}^V$. Note that all the dimensions of the $\mathbf{g}^A$, except the fourth, exhibit higher KLD values than the activation attribute, and at specific dimensions, KLD is almost 4 times higher than the KLD of the activation. A similar trend can be observed for the valence embedding vector $\mathbf{g}^V$, where five dimensions exhibit

higher KLD values than the valence attribute, and the largest KLD is extracted for the 6th dimension of the $\mathbf{g}^V$. These higher KLD values for the GRU based feature dimensions can be observed as the discriminative cues for the human-centric videos that also contribute to the performance of the proposed AVSUM architectures.

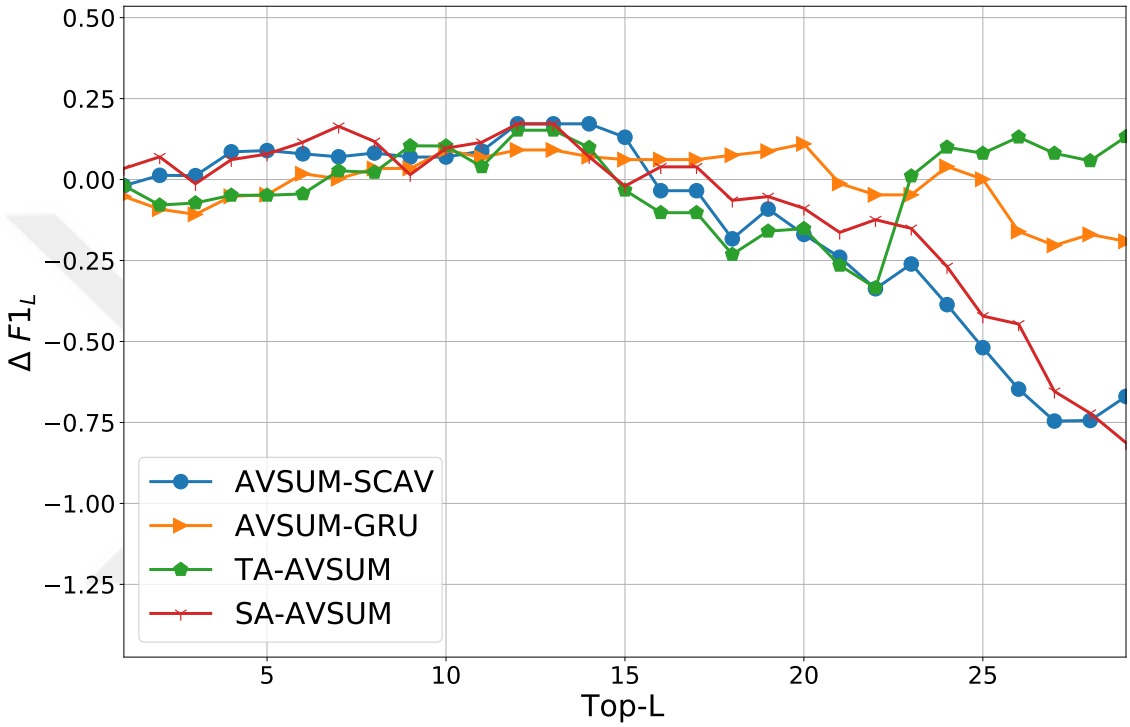


Figure 3.6: Performance comparison of the AVSUM models with the $\Delta F1_L$ metric for the *Top-30* human-centric videos at the TvSUM dataset

Video Level Evaluations

Figure 3.6 depicts the performance comparison of the AVSUM models with the $\Delta F1_L$ metric for the *Top-30* human-centric videos and yields valuable insights. Note that positive accumulation of $\Delta F1_L$ metric indicates a better performance than the baseline SUM-FCN model. In Figure 3.6, the AVSUM-SCAV, TA-AVSUM and SA-AVSUM models perform better than the baseline till the *Top-15* human-centric videos, whereas AVSUM-GRU model sustains a higher performance till the *Top-20* human-centric videos. Similar trends in the AVSUM-SCAV and SA-AVSUM $\Delta F1_L$

performances can be due to these architectures' common late fusion mechanism. As L increases, we can assume the human-centric characteristic of the videos is getting weaker. Hence, AVSUM-SCAV, SA-AVSUM, and AVSUM-GRU models perform better for the top human-centric videos, and their performances start degrading as human-centric characteristics get weaker.

We also investigate video level F-score performances for the proposed AVSUM models. Figure 3.7 presents scatter plot of the video level $F1$ scores on the left column and video level R scores on the right column, where the diagonal in each figure represents similar performances of the compared models. Furthermore, the *Top-15* human-centric videos are color coded in blue to observe their comparative performances.

Video level F-score performances of the AVSUM-GRU vs. SUM-FCN tend to cluster around the diagonal that makes these two models to be most similar in terms of F-score performance. Furthermore, majority of the *Top-15* human-centric videos are on or above the diagonal, which indicates a stronger performance for the human-centric videos. Unlike F1-score, the face recall performance comparison depicted by Figure 3.7b has a scattered behavior providing improvements to some videos while degrading others. However, it is seen that the majority of the *Top-15* human-centric videos are affected positively. This observation is in line with the $F1_{15}$ performance of AVSUM-GRU.

Video level F-score performances of the AVSUM-SCAV vs SUM-FCN have a higher deviation from the diagonal. However, almost all the *Top-15* human-centric videos are on or above the diagonal. This indicates a strong performance improvement for the human-centric videos. In terms of face recall on Figure 3.7d, the majority of the videos are accumulated around the diagonal, indicating that AVSUM-SCAV and SUM-FCN are most similar in terms of face recall.

Video level F-score performances of the TA-AVSUM and SA-AVSUM models also have a high deviation from the diagonal. However, like AVSUM-SCAV majority of the *Top-15* human-centric videos are on or above the diagonal for both. Similar to F-score, face recall comparisons of TA-AVSUM and SA-AVSUM have a scattered

behavior depicted by Figure 3.7f, and 3.7h. However, different than SA-AVSUM, scattered points accumulated on the positive side for TA-AVSUM, stating a major performance improvement which is inline with its best achieving R_{15} performance for the Max R criterion.

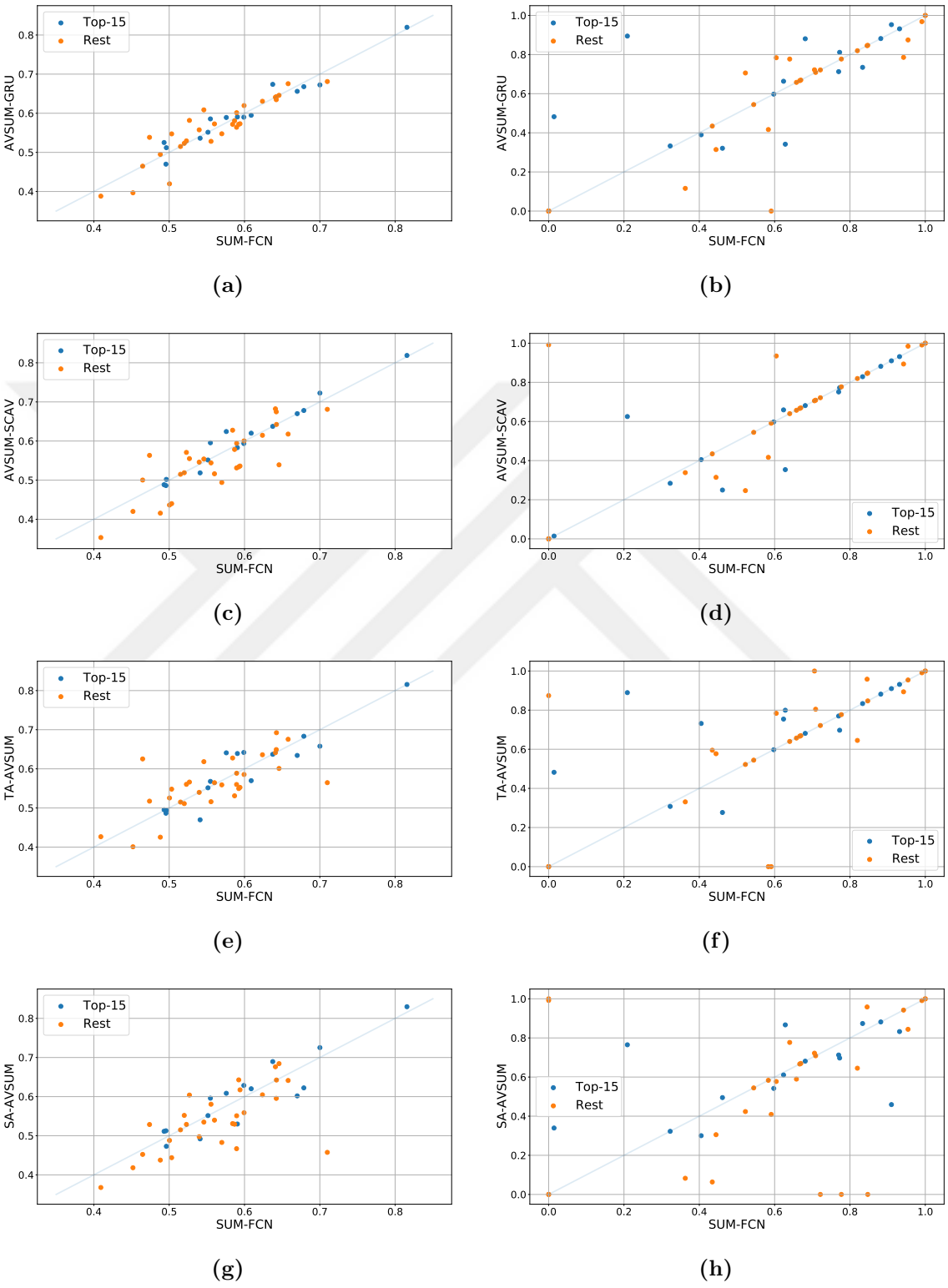


Figure 3.7: Video level $F1$ score and R score comparisons of the AVSUM models against the baseline SUM-FCN: $F1$ scores are with Max $F1$ criterion, R scores are with the Max R criterion, and the *Top-15* face including videos are color-coded in blue.

3.4 Conclusions on Affective Video Summarization

In this study, we proposed a new affective information enriched end-to-end video summarization framework for human-centric videos. As a first step, we modeled affective information in terms of AV attributes and GRU embeddings, which were extracted from the CER models. The CER-NET, a CER model achieving state-of-the-art CCC performance, was introduced. We explored the use of affective information with the proposed AVSUM-SCAV and AVSUM-GRU fusion models and attention mechanisms based TA-AVSUM and SA-AVSUM models. Experimental investigations of the proposed models were conducted on the RECOLA, TvSum and COGNIMUSE datasets.

We observed that with the fusion of affective information, F-score performance of the video summarization on the human-centric videos could be improved. To further analyze the effect of injected features, we defined a face recall (R) metric and showed that AVSUM-GRU and TA-AVSUM models outperform state-of-art models in face recall on both TvSUM and COGNIMUSE datasets. While AVSUM-GRU model has strong performance improvements for the human-centric videos on $F1$ score and face recall R, TA-AVSUM has significant performance improvements on face recall and is competitive on $F1$ as well. These two models are the most competitive against the baselines with the Max $F1$ criterion. On the other hand, we observed that attention-enhanced mechanisms exhibit strong performance gains with the Max R criterion for human-centric videos. We should also note that affective GRU embedding features exhibit higher KLD across with-face and without-face frames. Our evaluations on the predicted importance scores further verified strong performance of the TA-AVSUM model for the human-centric video summarization task. TA-AVSUM is introducing more than 15% improvement in rank correlation scores against SUM-FCN in both optimization criteria for human-centric videos while sustaining to stay in the top-two performing models with the rank correlation metrics.

Comparing the proposed AVSUM models, temporal attention-based TA-AVSUM

performs competitive with the Max $F1$ criterion and attains substantial improvement with the Max R criterion. Evaluations of the predicted importance scores further stress TA-AVSUM, as it introduces significant improvements regardless of the model selection criterion. The proposed AVSUM models integrate affective information to the summarization architectures and significantly improve video summarization for human-centric videos. Compilation of affective human-centric video datasets for video summarization tasks remains a critical and valuable future study. Investigation of multi-modal architectures for human-centric affective video summarization would be another valuable extension of the current work.

Chapter 4

AN AFFECTIVE VIDEO SUMMARIZATION DATASET

Recently with the increase in the diversity of smart devices (smartphones, smartwatches, gaming platforms, etc.) and the developments in cloud systems, a surge occurred in both users of intelligent devices and interaction duration. In addition, social media platforms (e.g. YouTube, TikTok, Twitch) that reward content generation increased their popularity; thus, the available online video material is witnessed exponential growth. The user content generation reinforcement has changed the plethora of video content shifted towards user-generated human-centric videos. The surge in the available video content and the shift towards human-centric content enforces efficient representation and video understanding centered on human behaviors. In parallel, consumer behavior, choosing a video to watch based on metadata, e.g., thumbnail, title, description, video length, requires automated solutions to generate metadata as the data become incomprehensible.

Conventional video summarization addresses representation and thumbnail generation problems by automatically generating a condensed version of a video, shorter by order of magnitude than the original, by selecting salient video segments or keyframes. While conventional video summarization is challenging, summarization of user-generated human-centric videos further increases the complexity of the conventional task as it involves complex human-to-human and human-to-computer communications. Conventional approaches extract a representation of each frame from the state-of-art image classification frameworks and build video summaries on top of these out-of-domain nonspecific features [Ji et al., 2020, Rochan et al., 2018], however representations are not representative enough to capture human behaviors. Recent works, [Tu et al., 2020] and [Köprü and Erzin, 2021] has addressed this issue by incorporating affective representations on top of the conventional visual features.

While these attempts are promising, [Tu et al., 2020] only considers video summary of human-centric videos as a post-processing step, and [Köprü and Erzin, 2021] injects prediction of emotion recognition network trained on another dataset, and thus, human-centric video summarization remains as a challenge.

Summarization of human-centric videos should address the problem of selecting salient video segments or keyframes and constructing an informative, interesting, and short summary of the human-centric actions in the video. As emotional processes drive cognitive (motor) processes, to summarize human-centric videos, the perceived emotion of the scene could be used to drive conventional video summarization. Then summarization of human-centric videos compromises extracting emotional state and selecting salient segments or keyframes. Joint optimization of these tasks requires a dataset involving human-centric videos with ground truth for emotion and video summarization.

In this work, we propose extending the AffWild2 database with video summarization annotations (AffWild2-VS) and provide a baseline for AVS on AffWild2-VS. The AffWild2 database contains annotations for arousal and valence for YouTube videos, and for the same videos, we have collected annotations for video summarization to bridge the gap for AVS. Unlike conventional annotation methods [Gygli et al., 2014, Potapov et al., 2014, Zlatintsi et al., 2017], by following the approach in [Song et al., 2015] we design annotation scenarios to avoid chronological bias (shots are scored higher if appearing early in a video) by pre-clustering and randomization, resulting in a high degree of inter-rater reliability.

To summarize, the main contributions of this study are as follows:

- We formulate end-to-end joint optimization for Affective Video Summarization (AVS) problem
- We introduce a novel dataset AffWild2-VS having annotations for AST and VS
- We further study the impact of audio-visual information on the annotations of video summarization

The rest of Chapter 4 organized as follows. Chapter 4.1 reviews the existing datasets in the CER and VS fields. Chapter 4.2 introduce data collection scenarios and dataset statistics. Chapter 4.3 presents predictive models studied on AffWild2-VS dataset. Chapter 4.4 presents experiments and discussion. Finally, Chapter 4.5 present an overall review.

4.1 Related Work

This study examines the summarization of human-centric videos and affective computing jointly and introduces a database annotated for both tasks. Hence, the availability of data and proof-of-claims brings a new perspective to the affective video summarization. In the following, we will introduce a few pieces of studies from both video summarization and affective computing fields related to our research.

Similar to our research, early video summarization works focus on genre-specific video summarization [Potapov et al., 2014] such as sports [Chen and De Vleeschouwer, 2011], egocentric [Sun et al., 2016, Lee et al., 2012] and news, and to produce summarizes they utilize genre-specific information: e.g. salient objects in sport. However, none of these genres dominates the video plethora, thus, lacks generalizability, unlike our case.

Table 4.1: A brief summary of the state-of-the-art databases

Database	Task	Video Summary Label	Duration	Human Centric Portion (%)	Video Style
SumME [Gygli et al., 2014]	VS	Binary	1h20mins	10	Natural
TvSUM [Song et al., 2015]	VS	Importance Scores	3h15mins	30	Natural
COGNIMUSE	VS, AST	Binary	3h30mins	85	Acted

Table 4.1 depicts state-of-the-art databases (TvSum [Song et al., 2015], SumMe[Gygli et al., 2014] and COGNIMUSE [Zlatintsi et al., 2017]) on video summarization focused on collecting videos from various genres. TvSum [Song et al., 2015] dataset contains 50 YouTube videos from 10 different categories: e.g sandwich making, beekeeping, and dog show; however it contains only 15 human-centric videos as analysis in [Köprü and Erzin, 2021] states. In total, TvSum includes 3 hours and 15 minutes of video content and shot-level importance score annotation (1-5) from 20 annotators. The SumMe [Gygli et al., 2014] contains 25 videos, ranging from 1–6 min, with 15–18 ground truth annotations, resulting in binary annotations. However, none-of-these studies are in line with the changing trend with the limited amount of human-centric videos, and also they are suitable for only VS. In fact, the most com-

parable state-of-the-art database to AffWild2-VS is COGNIMUSE, as it contains annotations for affective states, video summarization, and audio-visual events for a continuous half-hour segment of seven Hollywood movies. However, our study differs from COGNIMUSE regarding video summary annotation granularity, label consistency, and video content. Like SumMe, COGNIMUSE provides video summary annotation in binary form, which prevents the most vital evaluation, importance score correlation analysis, as stated in [Otani et al., 2019]. Their label consistency (Cronbach α) regarding video summary annotations is slightly lower than other widely used datasets. As the video content of COGNIMUSE is Hollywood movies, it doesn't include real emotions and human behaviors; rather, they are acted, and like other databases, these movies are not based on user-generated human-centric videos.

With the availability of the TvSum, SumMe, and COGNIMUSE, video summarization attracted researchers, as the supervised learning methods became applicable. Since then, it has been common to formulate video summarization as a sequence-to-sequence problem in a supervised learning setup [Rochan et al., 2018, Ji et al., 2020, Yuan et al., 2019, Li et al., 2021]. There are also successful unsupervised formulations for video summarization utilizing reinforcement learning [Zhou et al., 2018] or to minimize distribution differences between videos and their summarization via generative-discriminative networks [Mahasseni et al., 2017, Apostolidis et al., 2020].

Famous affective computing datasets mainly differ in terms of whether the annotation is categorical (happy, anger and etc.) as in IEMOCAP [Busso et al., 2008], or continuous (AVD) as in AffWild2 [Kollias and Zafeiriou, 2018], and [Ringeval et al., 2013]. However, none of these widely used affective computing databases includes saliency annotation or video summarization-related information.

Due to lack of data, the intersection between affective computing and video summarization was a neglected field, but there are several early attempts at it [Xu et al., 2018, Tu et al., 2020]. In [Xu et al., 2018], emotion recognition, emotion attribution, and emotion-oriented summarization are investigated using learned representations

from video and text in a multitask learning fashion. They define video summarization as a post-processing optimization step where keyframes are selected such that emotion attribution is maximized. In the second related study, a multitask model to jointly capture emotion recognition and attribution is proposed [Tu et al., 2020]. Similar to [Xu et al., 2018], video summarization is formulated as a post-processing optimization problem and solved using MINMAX dynamic programming. These studies differ from our work as they define video summarization as a post-processing step, and instead of annotating videos for saliency, these studies annotated the quality of their summarization outputs.

4.2 AffWild2-VS Database

AVS is an unfathomed domain as there is no available dataset. To bridge this gap and stress the importance of AVS, we collected a new dataset AffWild2-VS, a video oriented database multimodally annotated for video summarization.

4.2.1 Data

The video corpora of the AffWild2 data is collected from YouTube in [Kollias and Zafeiriou, 2018]. To capture a variety of stimuli, "reaction" and other emotion-related words from 2-D emotion wheel are used as the query term resulting in *movie or series reaction* videos, *joke or suprise* videos, and etc. The diversity of video types resulted in a database with a wide range of subjects and emotions.

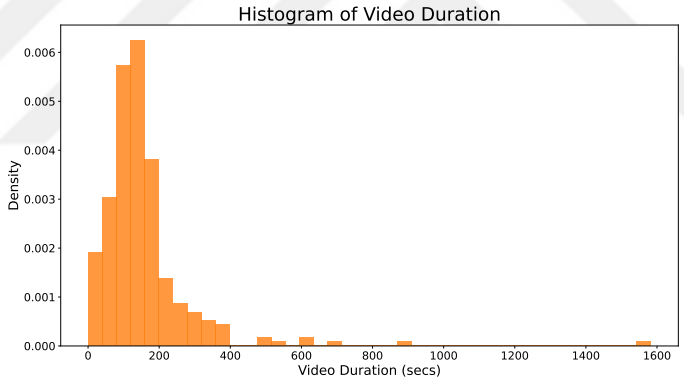


Figure 4.1: Distribution of video durations selected for video summarization task

From the AffWild2 videos, 291 of them chosen, corresponding to 12 hours of video to be annotated for video summarization, and Figure 4.1 depicts the duration distribution of selected videos. Figure 2 shows thumbnails of the 10 videos from each category.

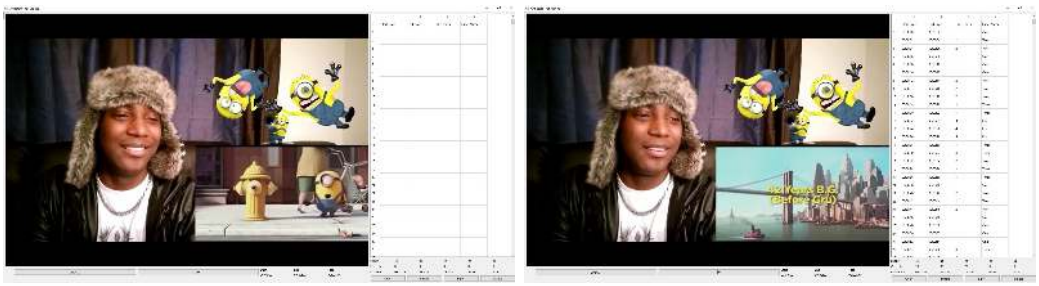
4.2.2 Annotation

It is infeasible to collect golden ground truths for video summarization due to the subjectivity of the task. To increase the reliability of the ground truths, we col-

laborated with Ango AI, a Turkish startup that specializes in the image and video annotation, where we hired professional annotators to rate the importance of the segments in the video and collected 2225 responses (5 annotators per video). Figure 3 depicts the annotation framework designed for video summarization annotation.

Annotation Scenario

As the VS task is subjective, there are a couple of annotation scenarios to generate ground truth. A goal-centric approach is to ask annotators to evaluate the quality of the summaries; however, this approach restricts initial summary generation to be either random or unsupervised. In addition, it requires the evaluation of summaries for each change in the generated summaries. By following [Song et al., 2015] and we ask annotators to follow: 1- Watch the video till the end, 2- Re-watch the video and evaluate the importance of the video segment by assigning the segment to scores from 1 (not important) to 5 (extremely important). The annotation framework that annotators used during the process is depicted in the Figure 4.2a.



(a) **(b)**

Figure 4.2: Annotation Framework: (a) before the annotation (b) during the annotation.

Importance Scores: Comparison of evaluation methodologies of VS in [Otani et al., 2019] stress the conventional classification metrics are insufficient, and rather correlation metrics between predicted and ground truth importance scores are appropriate for VS. As these correlation-based evaluations require non-binary labels,

Figure 4.3: Distribution on annotations without distribution requirements on small portion

thus unlike other VS databases[Zlatintsi et al., 2017, Gygli et al., 2014], we have collected importance scores to meet this rising demand by enabling five steps rather than binary values as ground truth.

Chronological bias: Similar to [Song et al., 2015], we observed that recency bias occurs in the annotation process, and annotators tend to assign higher scores to the shots that appear earlier in the video regardless of their audio-visual importance. To prevent chronological bias, after the original video is watched by the annotator, before the annotation phase each video is segmented into 30-sec shots and shuffled. Then annotators assign scores on the shuffled and merged video.

Regularizing score distributions: Since the video summarization task is subjective, there is no global standard of how long a summary should be or a golden ratio for summary and video length. However, when there is no instruction on the expected distribution of importance scores, annotators tend to skew the distribution towards the high values, as depicted in Figure 4.3. To alleviate this problem, following [Song et al., 2015] we instruct annotators to consider a target relative frequency ranges, which is to score 5 should be between 1% and 5%; score 4 should be between 5% to 10%; score 3 should be between 10% and 20%; score 2 should be between 20% and 40%, and score 1 gets the rest.

Another aspect of VS is some frameworks require binary labels, hence we convert 5 step importance scores as in the following:

$$s[n] = \begin{cases} 0 & \text{if } \zeta[n] < 3 \\ 1 & \text{if } \zeta[n] \geq 3, \end{cases} \tag{4.1}$$

where $\zeta[n]$ is 5 step importance score, and $s[n]$ corresponding binary label for the n^{th} frame.

4.2.3 Evaluation Metric

Motivated by [Otani et al., 2019], we assess the quality of ground truth by agreement between annotators in a leave-one-out fashion. This assessment than would be the golden standard for this dataset to compare predictive models with human-level performance. Precisely we calculated Kendall's Tau and F1-score across annotators.

Let ζ^k be the importance scores from k^{th} annotator, and ζ^l be the importance scores from l^{th} annotator where $k, l \in \{1, \ldots, A\}$ for the v^{th} video. Then $(\zeta_i^k, \hat{z}_i^k)$ and (ζ_j^k, ζ_j^l) are concordant pairs if either both $\zeta_i^k > \zeta_j^k$ and $\zeta_i^l > \zeta_j^l$ holds or both $\zeta_i^k < \zeta_j^k$ and $\zeta_i^l < \zeta_j^l$, else $(\zeta_i^k, \hat{z}_i^k)$ and (ζ_j^l, ζ_j^l) are considered as discordant pairs. Based on the concordant and discordant pairs, Kendall's τ is defined for the v^{th} video as:

$$\tau_v = \frac{1}{A}\frac{1}{A-1}\sum_{k=1}^{A}\sum_{l=1,k\neq l}^{A}\frac{N_c^{k,l} - N_d^{k,l}}{\sqrt{(N_c^{k,l} + N_d^{k,l} + T^{k,l})(N_c^{k,l} + N_d^{k,l} + U^{k,l})}}, \tag{4.2}$$

$$\tau = \frac{1}{V}\sum_{v=1}^{V}\tau_v \tag{4.3}$$

where $N_c^{k,l}$, $N_d^{k,l}$, $T^{k,l}$ and $U^{k,l}$ are the number of concordant pairs, discordant pairs, ties in ζ^k, and ties in ζ^l for the v-th video.

For the k-th annotator at v^{th} video, let the annotated binary summarization vector be $\mathbf{s}^k$ and corresponding ground truth vector to be $\mathbf{s}^l$ where $k, l \in \{1, \ldots, A\}$, $k \neq l$, and $\mathbf{s}^k, \mathbf{s}^l \in \mathbf{R}^N$ where A is number of annotators, and N is number of frames. Then precision P^k and recall R^k for the k-th annotator on v^{th} video are calculated

as

$$P_v^{k,l} = \frac{\mathbf{s}^k \cdot \mathbf{s}^l}{\sum_j^N s_j^l} \quad \text{and} \quad R_v^{k,l} = \frac{\mathbf{s}^k \cdot \mathbf{s}^l}{\sum_j^N s_j^k}, \tag{4.4}$$

where j runs over the frames. Video level F-score is defined as the harmonic mean of the precision and recall,

$$F1_v^k = \frac{1}{A-1} \sum_{l=1, l\neq k}^{A} \frac{2P^{k,l}R^{k,l}}{P^{k,l} + R^{k,l}}. \tag{4.5}$$

Then, the final F-score metric $F1$ is for video v, defined as the unweighted average of the pairwise annotator F-score values as

$$F1_v = \frac{1}{A} \sum_{k=1}^{A} F1_v^k, \tag{4.6}$$

$$F1 = \frac{1}{V} \sum_{v=1}^{V} F1_v \tag{4.7}$$

4.2.4 Statistics for Summarization Annotations

Figure 4.4, depicts the importance score distributions per annotator. It is seen that regarding scoring high importance scores (4-5) annotator 1, 2, 3, and 5 act similarly while *annotator 4* tend to assign higher scores frequently.

Table **??** depicts label consistency metrics with Cronbach α, F1, and Kendall's τ for two annotation scenarios *visual* and *audio-visual*. Visual only annotations achieve a Cronbach α of 0.88 while audio-visual annotations achieve 0.86. This behavior, higher consistency for visual annotations than audio-visual annotations, occurs in F1 and Kendall's τ where the former achieves 0.39 and 0.5, while latter achieves 0.38 and 0.47 respectively. Comparing consistency metrics of AffWild2-VS with state-of-art datasets TvSuM [Song et al., 2015], SumME [Gygli et al., 2014], and COGNIMUSE [Zlatintsi et al., 2017], our database show a higher degree of consistency.

Further Table 4.2 and Table 4.3 depicts pairwise F1 ($F1^{k,l}$) scores for visual and audio-visual scenarios. The slight discrepancy between *annotator-4* and the rest that is depicted in Figure 4.4, arises at the pairwise F1 scores as well. It is

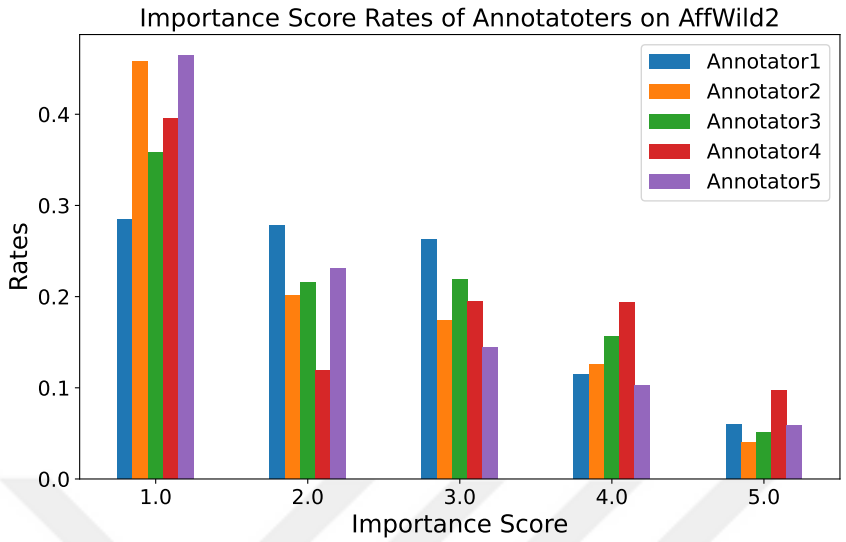


Figure 4.4: Distribution of importance scores per annotator

Table 4.2: Pairwise F1 ($F1^k$ (%)) comparison of annotators for visual only scenario

Annotator ID	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 4**	**Annotator 5**
Annotator 1	100	36.20	42.57	34.53	41.07
Annotator 2	-	100	41.56	33.35	44.65
Annotator 3	-	-	100	37.05	41.99
Annotator 4	-	-	-	100	35.52
Annotator 5	-	-	-	-	100

interesting that this discrepancy is higher for audio-visual scenario as the average F1 reduction from visual only to audio-visual scenario is 1.08 while for *annotator-4* the reduction is 1.45.

Table 4.3: Pairwise F1 ($F1^k$) comparison of annotators for audio-visual scenario

Annotator ID	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 4**	**Annotator 5**
Annotator 1	100	37.58	41.39	32.78	38.50
Annotator 2	-	100	38.34	34.09	48.60
Annotator 3	-	-	100	33.01	38.63
Annotator 4	-	-	-	100	34.73
Annotator 5	-	-	-	-	100

4.3 Affective Video Summarization Framework

We introduce our AVS framework with VS and AST modules. First, we will define AVS in two-fold, then provide an initial analysis of the relationship between emotion and video summarization. Finally, we will provide early predictive studies on the AffWild2-VS database.

4.3.1 Problem Statement

Following VS literature, we formulate AVS problem as a binary classification problem where the inputs are enriched with affective information. In this work, we approach AVS problem twofold: first as a vanilla type classification problem and as a sequence-to-sequence problem.

Our solution emits a single classification result for vanilla-type classification problems, whether to include the current frame into the summary or not, from a window of AV attributes. For the sequence-to-sequence formulation, our solution emits a sequence of binary values representing the evaluation result of the whole video.

Let ground truths for AV attributes be $s^{AV}[n]$ for the n^{th} frame, and $f^{AV}[n]$ be the AV attribute feature to be used for AVS. In this study, we feed ground truth affective attributes as the features, such that $f^{AV}[n] = S^{AV}[n]$. For the vanilla classification problem, we capture a window (w_s) of AV as a feature vector $\mathbf{f_n}$ for the classification of the n^{th} frame as:

$$\mathbf{f_n} = [\mathbf{f^{AV}[n - \Delta]}, \ldots, \mathbf{f^{AV}[n + \Delta]}] \ \in \mathbb{R}^{1\times 2\Delta\cdot 2}, \tag{4.8}$$

where Δ is windows size indicator $w_s = 2\Delta{+}1$.

For the sequence-to-sequence problem, the input to our models are matrices rather than vectors in the vanilla case. Let $\mathbf{f}^{AV}$ be the affective attribute matrix defining a video with N frames, then:

$$\mathbf{f}^{AV} = [\mathbf{f^{AV}[1]}, \ldots, \mathbf{f^{AV}[N]}] \ \in \mathbb{R}^{N\times 2}. \tag{4.9}$$

In addition let the visual features representing a video be $\mathbf{v}^v \in \mathbb{R}^{N \times \upsilon}$ where υ be number of features representing a frame. Then in sequence-to-sequence AVS we define three input types:

$$\mathbf{v}^v, \tag{4.10}$$

$$\mathbf{v}^{AV} = \mathbf{f}^{AV} \tag{4.11}$$

$$\mathbf{v}^{VAV} = \mathbf{v}^v \oplus \mathbf{f}^{AV}, \tag{4.12}$$

where $\oplus$ operator represents the feature vector concatenation per frame over a video.

4.3.2 Affective States and Video Summarization Relation

Our central claim in this thesis is salient regions of human-centric videos are related to the salient regions at the affective contour hence affective states should be correlated with the video summary annotations.

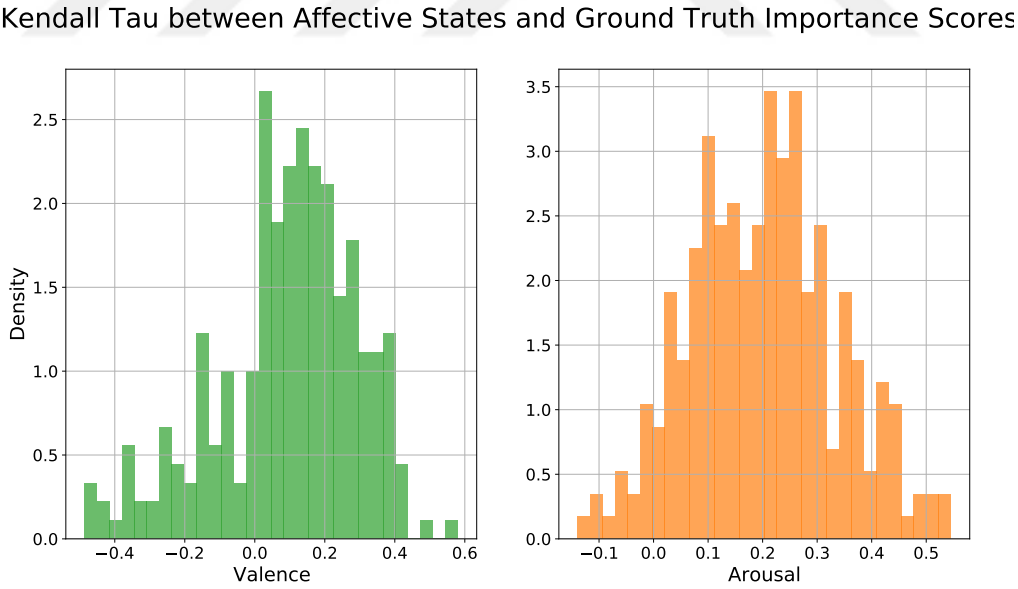


Figure 4.5: Distribution of Kendall Tau correlations between AV with Video Summarization ground truths

Figure 4.5 depicts the Kendall τ correlations between AV and VS importance scores amongst the AffWild2-VS. It is noticeable as the correlation distributions

are shifted towards to positive side with $N(0.2, 0.13)$ for arousal, and $N(0.09, 0.2)$. These results prove our claim on saliency detection via affective attributes.

4.3.3 Classifiers

The proposed AVS architectures are based on one of our previous study AVSUM [Köprü and Erzin, 2021] and a fully convolutional neural network (FCN) for semantic segmentation [Long et al., 2014] which is adapted for VS as SUM-FCN [Rochan et al., 2018].

SUM-FCN

SUM-FCN is based on 1-dimensional convolutional and deconvolution neural network structures as an encode-decoder.

Due to the nature of CNN, a fixed-sized input is required. Hence original videos are downsampled to a fixed number of frames imitating the fixed size cropping in image segmentation.

AVSUM

To understand how affective states can drive VS process, we modify SUM-FCN such that instead of operating on visual information, it operates on AV attributes and is referred to as AVSUM. Hence the input to the AVSUM becomes v^{AV}

Following [Köprü and Erzin, 2021] to drive the summarization process, we employ an early fusion. Hence the input to the AVSUM-AV is v^{VAV}.

Affective and Temporal Awareness Induced Fully Connected Summarization

In conventional setups, for instance, our vanilla-classification, fully connected neural networks lack learning temporal relationships and are problematic while modeling variable-sized sequences. In AVS setup, we overcome these shortages as videos are cropped to a fixed size, and we induce temporal awareness by passing the flattened f^{AV} as input and having an output size equal to the number of frames N. This

structure is equal to a classification problem with N classes, and we refer to this structure as ATFCN, which is depicted in Figure 4.6.

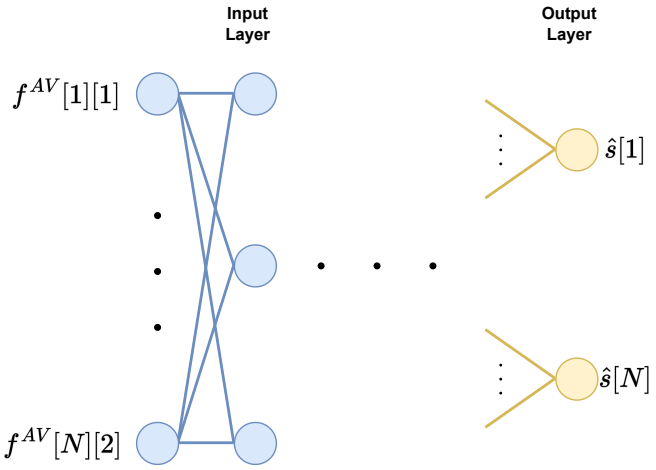


Figure 4.6: ATFCN Framework

4.3.4 Model Training

Since the core VS problem is defined based on a binary classification problem. Training of the AVS aforementioned models are similar and use cross-entropy loss. However, due to the imbalanced nature of the key frame selection problem, calculated losses are weighted to prevent imbalance. Let L_i be loss for i^{th} sample, then vanilla classification loss:

$$\mathcal{L}[i] = -w_1(z_i \log \hat{z}_i) + w_0((1 - z_i) \log 1 - \hat{z}_i), \tag{4.13}$$

$$\mathcal{L}_{vc} = \frac{1}{B} \sum_{i=1}^{B} L[i], \tag{4.14}$$

where B is the batch size, $z_i \in \{0, 1\}$ is the binary ground truth, $\hat{z}$ is the predicted score, and w_{z_i} is the weight of the i^{th} frame.

$$w_0 = \frac{1}{M} \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N} z_m[n], \text{ and } w_1 = 1 - w_0, \tag{4.15}$$

where $z_m[n]$ is the n^{th} sample from m^{th} video in the training dataset.

The loss function that is used to train AVSUM, AVSUM-AV, ATFCN, and SUM-FCN is similar to Equation 4.14, however, the sum is over the frames of a video instead of a batch as in the following:

$$\mathcal{L}_{avs}[i] = -w_{z_i}(z_i \log \hat{z}_i + (1 - z_i) \log 1 - \hat{z}_i), \tag{4.16}$$

$$\mathcal{L}avs = \frac{1}{N} \sum_{i=1}^{N} L_i. \tag{4.17}$$

Different than Equation 4.15, the weights for binary targets are defined as calculated per video:

$$w_0 = \frac{1}{N} \sum_{i=1}^{N} z_i \text{ and } w_1 = 1 - w_0 \tag{4.18}$$

4.4 AffWild2-VS Experimental Evaluation

Performances of proposed models are compared on the AffWild2-VS dataset using the metrics defined in Section 4.2. In this section, we will provide implementation details, and then early experimental results will be presented and discussed.

Table 4.4: Comparison of video summarization performance on AffWild2-VS

Model	**F1 (%)**	**R (%)**	τ
SUM-FCN [Rochan et al., 2018]	23.12	32.23	0.016
DRDSN [Zhou et al., 2018]	27	**38.1**	-
AVSUM	26.47	34.25	0.016
AVSUM-AV	26.93	34.41	0.016
TAFCN	**27.77**	35.18	**0.087**

Table 4.4 presents the F1, R, and Kendall's τ performances of AVS frameworks and state-of-the-art models. Comparing $F1$ performances of the proposed methods with SUM-FCSN [Rochan et al., 2018], proposed models outperform SUM-FCN with at least 3 %. Comparing Recall (**R**) performances, DRDSN outperforms other architectures, while TAFCN also achieves the second best R. To note, both AVSUM and AVSUM-AV outperform the baseline SUM-FCN.

Comparing Kendall's τ while convolutional architectures achieve similar performance, TAFCN outperforms others multiplying their performance by 6.

The performance results in Table 4.4 are in line with our findings, in Section 4.2, and when injected affective information enhances VS of human centric videos.

4.5 Conclusions on Affective Video Summarization Dataset

This study presented a multimodal human-centric video-oriented database annotated with uni-modal and multimodal sensory information for video summarization and introduced the AVSUM framework that uses affective information to summarize human-centric videos.

The content and subject of the videos drive the saliency; understanding and detecting the salient regions in a human-centric video requires understanding the complex relationships of human beings. Motivated by this fact, we collect VS summaries on a database having affective annotations. The purpose of AffWild2-VS is to enable joint research between human-behavior modeling, affective computing, video understanding, and saliency detection.

To prove our claims on the relationship between affective computing and video summarization, we conduct a correlation analysis on AffWild2-VS, stating arousal and valence attributes positively correlate with video summaries.

In addition, early predictive models are presented which use affective attributes to generate video summaries, and evaluation results state that AVSUM, AVSUM-AV, and especially TAFCN outperform state-of-the-art models on AffWild2-VS.

For future work, we would like to investigate representations of visual information in human-centric videos and their effect on video summarization. Moreover, we are interested in conducting multimodal-multitask studies for AVS.

Chapter 5

CONCLUSION

With the change in the video plethora, we identify affective video summarization, a novel challenging problem, at the intersection of video understanding, video summarization, and affective computing. This thesis attempts to demonstrate the significance of affective video summarization by exploring the different combinations of myriad angles in conjoint task while not significantly focusing on them.

In this final section, we will summarize the thesis by concentrating on our contributions.

5.1 Summary of Thesis

In this thesis, we proposed an affective video summarization framework starting with the question of whether does saliency of a human-centric video, and other videos (documentary, movie and etc) should be the same. Following the psychological studies claiming human behaviors are driven by emotional processes, unlike other beings, we claim that salient regions of a human-centric video should be correlated with the salient regions of affective contour. To achieve an affective video summarization framework, we investigate affective computing in Chapter 2. We investigate fusion methods to drive video summarization with affective information in Chapter 3, then considering the shortcomings in Chapter3 which is the lack of data for annotated for both affective computing and video summarization, we introduce AffWild2-VS in Chapter4 and investigate video summarization via affective states.

In Chapter 2, we addressed the following questions: how can we effectively track affective contour and capture the most salient regions in the affective contour? For the former, we defined CER as a regression of AVD attributes; for the latter, we define ABS and binary classification for ABS detection. Noting the correlations

between AVD attributes, we propose MTL-CER a network trained to estimate AVD attributes jointly using multi-task learning. In addition, we showed that standard loss functions lacked performance when CER solutions were evaluated with CCC and proposed correlation-based losses, which outperformed standard regression losses. Our investigation of audio-visual modalities' contribution shows that multi-modal architectures outperform unimodal architectures. On the other hand, while CER modules trained with correlation-based losses improve both performance and the overall quality of experience, they lack in detecting instantaneous changes. We defined ABS as a mathematical formulation that captures the affective contour's rapid changes. We proposed dilated kernel to increase temporal receptive field and kernel fusion which improves the detection of affective bursts by 3 pp and 2 pp, respectively.

Subsequently, in Chapter 3, we addressed how we can incorporate affective and visual information to capture salient regions in the human-centric video. Noting the lack of annotated dataset for both affective computing and video summarization, we proposed CER-NET mimicking MTL-CER in Chapter 2 to extract affect related information. We investigated the fusion of predicted AV attributes and GRU embeddings to a commodity video-summarization framework. Investigation of fusion mechanism stated that attention-based fusion at the latent space is more sensitive to capturing human-centric contents in a mixed video database. To demonstrate the effectiveness of our models, considering the state-of-art databases lack human content, we defined new concepts and metrics; Max-R, Top-L, ΔF1. In addition, we showed that AVS models are also competitive in terms of state-of-the-art evaluation metrics, and especially TA-AVSUM outperforms state-of-art models on the Kendall-τ and Spearman's Rank Correlation coefficient.

In Chapter4, we introduced a new dataset, AffWild2-VS, to address the limitations encountered in Chapter 3. AffWild2-VS is a novel dataset with annotations for AV attributes and VS annotations collected in this study. We designed specific annotation scenarios to overcome human bias, chronological bias, and binarization and investigated the contribution of audio and visual modalities to saliency detec-

tion. Analysis of the annotations showed that the collected database has a higher label consistency than famous video-summarization datasets. In addition, with the availability of ground truth of AV attributes and VS, statistical analysis showed that AV attributes are correlated with the video summaries annotations. At the same time, arousal has a higher correlation than valence with the video summaries. Then we presented the AVS framework as an early attempt to generate video summaries from AV attributes; experimental studies showed that our proposition outperforms state-of-the-art video summarization frameworks.

To sum up, this thesis broadly focuses on a new problem affective video summarization which is in the intersection of video understanding, representation and affective computing. We aim to attract attention from these fields by showing the importance of the problem and providing data to be studied. All these works encourage us to deeper studies on AVS problem and AffWild2-VS dataset, and we will close this thesis by identifying several directions that can be further investigated.

5.2 Future Directions

In this section, we outline the aspects of the future work, based our intutions gained from experimental studies.

- **Multimodality** We showed that multi-modality helps to improve performance of the frameworks in CER problem. Audio-visual multi-modal investigations can be conducted on the affective burst detection problem, and various fusion techniques can be surveyed to enhance the performance. Regarding video summarization, similarly, multi-modal architectures can be investigated by considering audio, video, and text to enhance the performance.

- **Multi-task Learning** With the availability of AffWild2-VS, multi-task learning-based frameworks where affective computing model and video summarization model is trained jointly can be investigated.

- **Explainable Video Summarization and Affective Computing** Although

we try to explain which type of affective information is more feasible and why is more feasible in AVS concept in the mixed video distribution contents. The Explainability of saliency detection on audio-visual remain remains an interesting challenge.

- **Unsupervised approaches** Unsupervised approaches especially based on generative-discriminative frameworks, can be investigated on AVS jointly or separately for affective computing and video summarization.

5.3 List of Publications

A list of publications resulting from this thesis is shown as follows:

Publications

- Köprü, Berkay, and Engin Erzin. "Use of Affective Visual Information for Summarization of Human-Centric Videos." In IEEE Transactions on Affective Computing, under review.

- Köprü, Berkay, and Engin Erzin. "Multimodal continuous emotion recognition using deep multi-task learning with correlation loss." arXiv preprint arXiv:2011.00876 (2020).

- Köprü, Berkay, and Engin Erzin. "Affective Burst Detection from Speech using Kernel-fusion Dilated Convolutional Neural Networks." In 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022.

BIBLIOGRAPHY

[Anagnostopoulos et al., 2015] Anagnostopoulos, C.-N., Iliou, T., and Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177.

[Apostolidis et al., 2020] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2020). Unsupervised video summarization via attention-driven adversarial learning. In *MultiMedia Modeling*, pages 492–504. Springer International Publishing.

[Apostolidis et al., 2021] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2021). Video summarization using deep neural networks: A survey. *arXiv:2101.06072v1*.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR 2015, 3rd International Conference on Learning Representations*.

[Baxter, 1997] Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.

[Bhattacharya et al., 2021] Bhattacharya, U., Wu, G., Petrangeli, S., Swaminathan, V., and Manocha, D. (2021). HighlightMe: Detecting Highlights from Human-Centric Videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8137–8147.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive

emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

[Castellano, 2008] Castellano, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction*, pages 92–103. Springer.

[Castellano et al., 2008] Castellano, G., Kessous, L., and Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction*, pages 92–103. Springer.

[Chen and De Vleeschouwer, 2011] Chen, F. and De Vleeschouwer, C. (2011). Formulating team-sport video summarization as a resource allocation problem. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(2):193–205.

[Chen et al., 2017] Chen, S., Zhao, J., Jin, Q., and Wang, S. (2017). Multimodal multi-task learning for dimensional and continuous emotion recognition. *AVEC 2017 - Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, co-located with MM 2017*, (October):19–26.

[Ekman and Davidson, 1994] Ekman, P. and Davidson, R. J. (1994). *The Nature of Emotion: Fundamental Questions*. Oxford University Press, New York.

[Eyben et al., 2016] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

[Fajtl et al., 2018] Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., and Remagnino, P. (2018). Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer.

[Fatima and Erzin, 2017] Fatima, S. N. and Erzin, E. (2017). Cross-subject continuous emotion recognition using speech and body motion in dyadic interactions. In *INTERSPEECH 2017*, pages 1731–1735.

[Graves and Jaitly, 2014] Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772.

[Gunes and Pantic, 2010] Gunes, H. and Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99.

[Gygli et al., 2014] Gygli, M., Grabner, H., Riemenschneider, H., and Gool, L. V. (2014). Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.

[Han et al., 2017] Han, J., Zhang, Z., Ringeval, F., and Schuller, B. (2017). Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5005–5009.

[Han et al., 2014] Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association.*

[Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *CVPR 2018, IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

[Huang et al., 2015] Huang, Z., Epps, J., and Ambikairajah, E. (2015). An Investigation of Emotion Change Detection from Speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Ji et al., 2020] Ji, Z., Xiong, K., Pang, Y., and Li, X. (2020). Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717.

[Joho et al., 2011] Joho, H., Staiano, J., Sebe, N., and Jose, J. M. (2011). Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, 51(2):505–523.

[Jung et al., 2020] Jung, Y., Cho, D., Woo, S., and Kweon, I. S. (2020). Global-and-local relative position embedding for unsupervised video summarization. In *European Conference on Computer Vision*, pages 167–183. Springer.

[Khorram et al., 2017] Khorram, S., Aldeneh, Z., Dimitriadis, D., McInnis, M., and Provost, E. M. (2017). Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. *arXiv preprint arXiv:1708.07050*.

[Khorram et al., 2018] Khorram, S., Jaiswal, M., Gideon, J., McInnis, M., and Provost, E. M. (2018). The priori emotion dataset: Linking mood to emotion detected in-the-wild. *arXiv preprint arXiv:1806.10658*.

[King, 2009] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758.

[Kojouharova et al., 2019] Kojouharova, P., File, D., Sulykos, I., and Czigler, I. (2019). Visual mismatch negativity and stimulus-specific adaptation: the role of stimulus complexity. *Experimental Brain Research*, 237(5):1179–1194.

[Kollias and Zafeiriou, 2018] Kollias, D. and Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*.

[Köprü and Erzin, 2020] Köprü, B. and Erzin, E. (2020). Multimodal continuous emotion recognition using deep multi-task learning with correlation loss. *arXiv preprint arXiv:2011.00876*.

[Köprü and Erzin, 2021] Köprü, B. and Erzin, E. (2021). Use of affective visual information for summarization of human-centric videos. *arXiv preprint arXiv:2107.03783*.

[Koutras et al., 2015] Koutras, P., Zlatintsi, A., Iosif, E., Katsamanis, A., Maragos, P., and Potamianos, A. (2015). Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *2015 IEEE international conference on image processing (ICIP)*, pages 4361–4365. IEEE.

[Kovarski et al., 2017] Kovarski, K., Latinus, M., Charpentier, J., Cléry, H., Roux, S., Houy-Durand, E., Saby, A., Bonnet-Brilhault, F., Batty, M., and Gomot, M. (2017). Facial Expression Related vMMN: Disentangling Emotional from Neutral Change Detection. *Frontiers in Human Neuroscience*, 11:18.

[Kwon et al., 2003] Kwon, O. W., Chan, K., Hao, J., and Lee, T. W. (2003). Emotion recognition by speech signals. *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, pages 125–128.

[Lee et al., 2012] Lee, Y. J., Ghosh, J., and Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE.

[Li et al., 2021] Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., and Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 111:107677.

[Long et al., 2014] Long, J., Shelhamer, E., and Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651.

[Ma et al., 2005] Ma, Y.-F., Hua, X.-S., Lu, L., and Zhang, H.-J. (2005). A generic framework of user attention model and its application in video summarization. *IEEE transactions on multimedia*, 7(5):907–919.

[Ma et al., 2002] Ma, Y.-F., Lu, L., Zhang, H.-J., and Li, M. (2002). A user attention model for video summarization. In *MULTIMEDIA '02, Tenth ACM International Conference on Multimedia*, pages 533–542.

[Mahasseni et al., 2017] Mahasseni, B., Lam, M., and Todorovic, S. (2017). Unsupervised video summarization with adversarial LSTM networks. In *CVPR 2017, 30th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2982–2991.

[Male et al., 2020] Male, A. G., O'Shea, R. P., Schröger, E., Müller, D., Roeber, U., and Widmann, A. (2020). The quest for the genuine visual mismatch negativity (vMMN): Event-related potential indications of deviance detection for low-level visual features. *Psychophysiology*, 57(6):e13576.

[Mehrabian, 1996] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

[Mehrabian, 2017] Mehrabian, A. (2017). Communication without words. In *Communication theory*, pages 193–200. Routledge.

[Metallinou et al., 2010] Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., and Narayanan, S. (2010). The usc creativeit database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, page 55.

[Mirsamadi et al., 2017] Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231.

[Money and Agius, 2009] Money, A. G. and Agius, H. (2009). Analysing user physiological responses for affective video summarisation. *Displays*, 30(2):59–70.

[Ng et al., 2015] Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM.

[Nicholson et al., 1999] Nicholson, J., Takahashi, K., and Nakatsu, R. (1999). Emotion recognition in speech using neural networks. *ICONIP 1999, 6th International Conference on Neural Information Processing - Proceedings*, 2:495–501.

[Otani et al., 2019] Otani, M., Nakashima, Y., Rahtu, E., and Heikkila, J. (2019). Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7596–7604.

[Parthasarathy and Busso, 2017] Parthasarathy, S. and Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-task learning. In *INTERSPEECH 2017*, pages 1103–1107.

[Parthasarathy and Busso, 2018] Parthasarathy, S. and Busso, C. (2018). Predicting emotionally salient regions using qualitative agreement of deep neural network regressors. *IEEE Transactions on Affective Computing*, 12(2):402–416.

[Paul, 1999] Paul, E. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

[Potapov et al., 2014] Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. (2014). Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer.

[Ringeval et al., 2018] Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., et al. (2018). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 3–13.

[Ringeval et al., 2015] Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015). Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *FG 2013, 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*.

[Rochan et al., 2018] Rochan, M., Ye, L., and Wang, Y. (2018). Video summarization using fully convolutional sequence networks. In *ECCV 2018, European Conference on Computer Vision*, pages 358–374.

[Russell and Mehrabian, 1977] Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

[Schlosberg, 1954] Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81.

[Schmitt et al., 2019] Schmitt, M., Cummins, N., and Schuller, B. W. (2019). Continuous Emotion Recognition in Speech — Do We Need Recurrence? In *INTERSPEECH 2019*, pages 2808–2812.

[Song et al., 2015] Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. (2015). TVSum: Summarizing web videos using titles. In *CVPR 2015, IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187.

[Sun et al., 2016] Sun, M., Farhadi, A., Taskar, B., and Seitz, S. (2016). Summarizing Unconstrained Videos using Salient Montages. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2256–2269.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR 2015, IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

[Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204.

[Tu et al., 2020] Tu, G., Fu, Y., Li, B., Gao, J., Jiang, Y. G., and Xue, X. (2020). A Multi-Task Neural Approach for Emotion Attribution, Classification, and Summarization. *IEEE Transactions on Multimedia*, 22(1):148–159.

[Tzirakis et al., 2021] Tzirakis, P., Chen, J., Zafeiriou, S., and Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53.

[Tzirakis et al., 2017] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using

deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

[Tzirakis et al., 2018] Tzirakis, P., Zafeiriou, S., and Schuller, B. W. (2018). End2you–the imperial toolkit for multimodal profiling by end-to-end learning. *arXiv preprint arXiv:1802.01115*.

[Valstar et al., 2016] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

[Vicol et al., 2018] Vicol, P., Tapaswi, M., Castrejon, L., and Fidler, S. (2018). Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590.

[Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[Wang et al., 2021] Wang, C., Zhang, J., Jiang, W., and Wang, S. (2021). A Deep Multimodal Model for Predicting Affective Responses Evoked by Movies Based on Shot Segmentation. *Security and Communication Networks*, 2021:1–12.

[Wang et al., 2022] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., and Zhang, W. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83-84:19–52.

[Wu et al., 2019] Wu, S., Du, Z., Li, W., Huang, D., and Wang, Y. (2019). Continuous emotion recognition in videos by fusing facial expression, head pose and

eye gaze. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 40–48, New York, NY, USA. Association for Computing Machinery.

[Xu et al., 2018] Xu, B., Fu, Y., Jiang, Y. G., Li, B., and Sigal, L. (2018). Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

[Yuan et al., 2019] Yuan, Y., Li, H., and Wang, Q. (2019). Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, 7:64676–64685.

[Zhang et al., 2016] Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. (2016). Video summarization with long short-term memory. In *ECCV 2016, European Conference on Computer Vision*, pages 766–782. Springer International Publishing.

[Zhang et al., 2018] Zhang, Z., Han, J., Coutinho, E., and Schuller, B. (2018). Dynamic difficulty awareness training for continuous emotion prediction. *IEEE Transactions on Multimedia*, 21(5):1289–1301.

[Zhang et al., 2019] Zhang, Z., Wu, B., and Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709.

[Zhou et al., 2018] Zhou, K., Qiao, Y., and Xiang, T. (2018). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI Conference on Artificial Intelligence*, volume 32.

[Zlatintsi et al., 2017] Zlatintsi, A., Koutras, P., Evangelopoulos, G., Malandrakis, N., Efthymiou, N., Pastra, K., Potamianos, A., and Maragos, P. (2017). 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):54.