ViPER: Video-based Perceiver for Emotion Recognition

ViPER: Video-based Perceiver for Emotion Recognition

Lorenzo Vaiani*
lorenzo.vaiani@polito.it
Politecnico di Torino
Turin, Italy

Moreno La Quatra*
moreno.laquatra@polito.it
Politecnico di Torino
Turin, Italy

Luca Cagliero
luca.cagliero@polito.it
Politecnico di Torino
Turin, Italy

Paolo Garza
paolo.garza@polito.it
Politecnico di Torino
Turin, Italy

ABSTRACT

Recognizing human emotions from videos requires a deep understanding of the underlying multimodal sources, including images, audio, and text. Since the input data sources are highly variable across different modality combinations, leveraging multiple modalities often requires ad hoc fusion networks. To predict the emotional arousal of a person reacting to a given video clip we present ViPER, a multimodal architecture leveraging a modality-agnostic transformer-based model to combine video frames, audio recordings, and textual annotations. Specifically, it relies on a modality-agnostic late fusion network which makes ViPER easily adaptable to different modalities. The experiments carried out on the Hume-Reaction datasets of the MuSe-Reaction challenge confirm the effectiveness of the proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**.

KEYWORDS

Video processing, emotion recognition, multimodal learning, modality-agnostic learning

1 INTRODUCTION

Human emotions are often conveyed through different channels and modalities, e.g., facial expressions, body language, voice. The ability to express emotions allows human beings to effectively interact with each other. For example, facial expressions convey crucial information in non-verbal communication [35].

Automating the process of emotion recognition from videos is known to be particularly relevant to various application contexts such as human behavior analysis [25], affective education [4], and healthcare [21]. However, in many real-world scenarios applying image processing to identify facial expressions is not sufficient to effectively recognize human emotions. The main reason is that facial expressions can be either ambiguous or dependent on the surrounding context. For instance, people can smile not only when they are happy but also when they are nervous, angry, or in front of an authority. Hence, there is often a need to effectively and efficiently combine multimodal data sources including images, audio recording, and the textual annotations [29].

This paper addresses the problem of automatic emotion recognition from videos using multimodal data representations. Existing approaches (e.g., [28, 29]) rely on fusion networks to combine data of multiple modalities. However, these networks are typically modality-dependent thus not easily adaptable to different scenarios and tasks.

We present ViPER, a new multimodal architecture to extract human emotions from videos. ViPER is designed to address the MuSe-Reaction challenge [8], which aims at automatically recognizing the emotional state of people reacting to a given video clip. More specifically, it entails estimating the intensity of seven different emotional reactions, i.e., adoration, amusement, anxiety, disgust, empathic pain, fear, and surprise. Video recordings of people reactions to a video clip are captured by a front-facing camera showing the person's face and recording her/his voice while watching the video.

ViPER relies on a Deep Learning architecture that combines multimodal data sources (e.g., images, audio, text) using an established modality-agnostic late fusion strategy. The facial expression, the audio recording, and the textual annotations associated with the video frames are jointly exploited to accurately predict the human reactions. The use of a modality-agnostic late fusion step, based on Perceiver [13], preserves the generality and portability of the proposed solution towards different scenarios. The proposed solution ensures model scalability by using attention bottlenecks that transform the inputs into fixed-length latent representations. The resulting embeddings have an independent computational cost regardless of the original input size.

The main contributions of this work can be summarized as follows:

- ViPER adopts a modality-agnostic fusion network that makes the proposed architecture adaptable to other input modalities. To the best of our knowledge, this is the first attempt to address modality-agnostic emotion recognition from videos.
- ViPER leverages the CLIP pre-trained model [23] to enrich the video frames with new textual annotations. The purpose is to augment the input sources with the textual descriptions of the video frames providing relevant insights into the emotional arousal of the person involved.
- We perform an ablation study of the different modality combinations. The fusion of images, audio, and text achieves the best performance across all the tested modality combinations.

*Both authors contributed equally to this research.

The project source code is available for research purposes only[1]. The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the ViPER architecture. Sections 4 and 5 describe the experimental design and summarize the main empirical outcomes. Finally, Section 6 draws the paper conclusions and future research directions.

2 PRIOR WORKS ON EMOTION RECOGNITION

Previous studies already addressed automatic emotion recognition from images, text, or a combination of data modalities. Hereafter, we will separately analyze the existing methods based on image processing, natural language processing, and multimodal learning. Furthermore, we also introduce the modality-agnostic learning context.

2.1 Approaches based on image processing

The recognition of human emotions from single image frames has been addressed using Convolutional Neural Networks (CNNs), which are established for extracting features from images [2, 18]. To contextualize the facial expressions some previous works also analyze the facial landmarks [30] and the Facial Action Units (FAUs) [20]. Recently, Transformer-based models [10] have shown to be very successful in various computer vision tasks, including facial expression recognition [16]. Despite their ability to attend relevant image portions, they are suboptimal for emotion recognition from videos because they neglect the temporal evolution of the detected patterns. To tackle the above-mentioned issue, Recurrent Neural Networks (RNNs) [11] and 3D Convolutional Neural Networks (3D-CNNs) [12] have been exploited to analyze sequences of video frames.

2.2 Approaches based on natural language understanding

Detecting emotions from natural language is a long-standing Natural Language Understanding (NLU) problem, which is commonly formulated as a text classification task [5]. Beyond voice transcriptions and human-generated video annotations, existing NLU methods also consider the main acoustic properties of the audio signals such as pitch and loudness [24]. Combining textual and acoustic features for emotion recognition has recently proved to be particularly promising [31]. Specifically, the authors leverage both temporal and semantic relationships between textual and acoustic features to perform high-quality emotion predictions. Despite non-verbal acoustic features play an essential role in human communication, their exploration to address emotion recognition from videos is still open [26]. This is the main purpose of the MuSe-Reaction challenge [8], which is addressed by ViPER.

2.3 Multimodal approaches

Multimodal emotion recognition entails combining data sources of different modalities to recognize the human emotions. They rely on fusion techniques aimed at jointly analyzing multiple modalities to capture complex dynamics of human emotions [34]. For instance, deep learning models (e.g., [19, 32]) can effectively combine the voice sound and the facial expression to improve the prediction accuracy.

The Multimodal Sentiment analysis challenge (i.e., MuSe) aims at fostering research in multimodal emotion recognition. Within this research contest, prior works already tried to use the attention mechanism to attend relevant portions of the input data [28, 29]. However, existing approaches to emotion recognition rely on modality-specific fusion networks. For example, they attend relevant regions in visual [10] and audio [7] data. As a drawback, modality-specific fusion networks are not easily adaptable to different modality combinations.

2.4 Modality-agnostic learning

Modality-agnostic architectures aim at efficiently learning complex data representation regardless of the modalities involved [13]. These architectures have been successfully applied to solve different tasks, including vision-language classification [3], image-text retrieval and visual question answering [36]. This paper investigates the use of a state-of-the-art modality-agnostic architecture in emotion recognition from videos. To the best of our knowledge, this work is the first attempt in this line of research.

3 METHODOLOGY

We present **ViPER** (namely, **Vi**deo-based **P**erceiver for **E**motion **R**ecognition), a multimodal architecture for emotion recognition from videos. ViPER is suited to address the MuSe-Reaction challenge [8], which aims at predicting the emotional arousal of a person watching a video clip. The key aspects addressed by MuSe-Reactions are

- The *multimodal* nature of the input data, which encompass both visual and acoustic features.
- The primary importance of *facial expressions* to solve the emotion recognition task.
- The prevalence of *non-verbal communication* in the audio recordings.

A sketch of the ViPER architecture is shown in Figure 1. To tackle the MuSe-Reactions task, ViPER adopts an attention-based, modality-agnostic late fusion strategy. The fusion network takes as input the *visual component*, i.e., the images corresponding to the video frames, and the *acoustic component*, i.e., the audio recording associated with the video. The visual component is exploited to elicit various kinds of features including transformer-based visual embeddings, Facial Action Units (FAUs), and frame captions. In particular, frame captioning allows us to augment the input data with a new modality consisting of a textual description of the video frames encoded using a state-of-the-art contextualized embedding model.

The original and augmented data produce the following latent features:

- **Visual features**: Vision Transformer (ViT) [10], Facial Action Units (FAUs) [20] (i.e., the yellow tokens in Figure 1).
- **Textual features (augmented)**: the RoBERTa contextualized embeddings [15] of the frame captions (i.e., the red tokens in Figure 1).

[1] https://github.com/VaianiLorenzo/ViPER

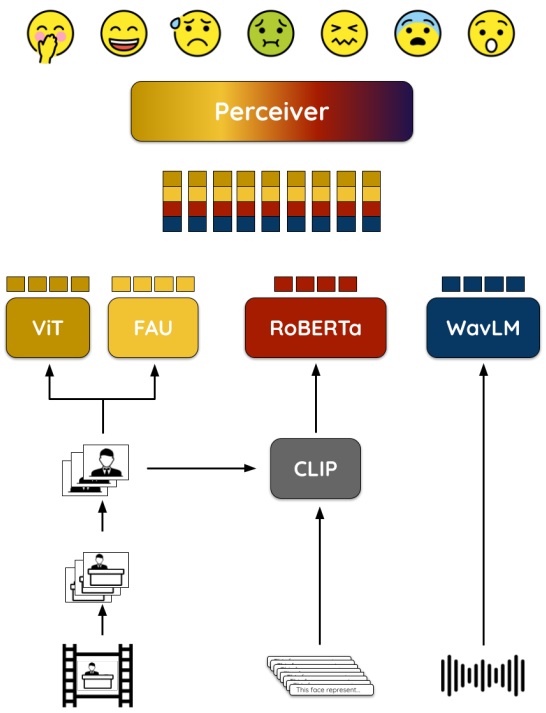

Figure 1: Sketch of ViPER architecture[2]

- **Acoustic features**: the x-vectors representations of the audio waveforms derived from WavLM [7] (i.e., the blue tokens in Figure 1).

Visual, textual, and acoustic features are combined together using a late fusion network, namely Perceiver [13]. Since it relies on a modality-agnostic approach, the network is inherently adaptable to different combinations of modalities.

A more detailed description of the ViPER components follows.

3.1 Image embeddings

Video data samples can be seen as a sequence of hundreds of images. Two consecutive images in the video sequence are likely to be highly similar to each other. Hence, we sample images and extract visual features from the sampled data. The per-video sampling frequency is adapted to its duration to avoid introducing imbalances in the sample sizes.

Since Transformer architectures [33] have proven their superiority in the computer vision domain, outperforming Convolutional Neural Networks, we use a Vision Transformer (ViT) [10] to extract the information conveyed by each selected frame. ViT has a BERT-like architecture and takes as input a 224×224 image. The image is divided into 16×16 patches, which are provided as input tokens to the transformer. Finally, we extract the image embeddings by considering the 768-dimensional last hidden state of a special *Classification Token* that is added to the input sequence. It encodes the global information of the image and is used as a representation of the frame for the next stage.

Table 1: Examples of template captions for MuSe-React emotion classes.

Emotion	Template caption
Adoration	This face is feeling adoration
Amusement	It looks like this face is amused
Anxiety	Anxiety is visible on this face
Disgust	The expression on this face is disgusted
Emphatic Pain	This face is experiencing Empathic Pain
Fear	Fear is apparent on this face
Surprise	The face on this picture is feeling surprised

3.2 Audio embeddings

To couple the visual information extracted from each frame with the acoustic information present in the video audio track, each audio recording is divided into N small fragments. Each of them has the same length, no overlap with the other ones and is centered in the timestamp corresponding to a selected video frame, including the first and last ones. The duration of the audio fragments associated with the video extremities is half of those of the other fragments. They respectively begin and end in correspondence with the associated frame.

To encode the content of each audio fragment, we use WavLM [7] model. It uses self-attention to extract acoustic representations from the raw waveforms. Originally proposed for the speech recognition tasks, WavLM has already demonstrated generalization capabilities for other tasks, including intent detection and speaker verification. Our approach applies the WavLM model to obtain feature-rich representations using the same scheme adopted by the x-vector technique. X-vectors [27] are representations derived from the raw waveforms, which are commonly used to extract speaker-related acoustic features from audio recordings. The corresponding embedding vectors are fed to a classification layer to learn the relationship between acoustic features and audio categories (e.g., speaker's identity or emotions). In such a way, we use the aforesaid audio encodings to generate acoustic representations for the input audio fragments.

3.3 Data augmentation based on frame captioning

We augment the visual features associated with each video frame with textual captions. To this end, we first generate 53 caption templates consisting of 6 to 9 descriptive sentences per target emotion. In addition to the emotions involved in the task, caption templates also cover the neutral emotion, which can be mapped to the video frames preceding or following the actual human reaction. Next, we match each video frame to the most pertinent caption template to obtain the corresponding textual description. To this aim, we exploit a CLIP encoder fine-tuned on image emotion recognition [6] to encode sentences and video frames into a common embedding space. Then, we compute the cosine similarity between the sentence and frame embeddings and use the sentence with the highest similarity

[2] All emojis designed by OpenMoji – the open-source emoji and icon project. License: CC BY-SA 4.0

score as the textual description of the frame. The corresponding sentence is then encoded using a pre-trained RoBERTa model [15] and the resulting 768-dimensional vector is concatenated with the other frame features to compose the input of our proposed model.

Table 1 reports some qualitative examples of template-based description associated with each emotion class. The complete list of template sentences is available in the project repository[3].

3.4 Facial Action Units

FAUs describe the activity of facial muscles that cause the movement of many elements of the human face, such as the eyes, lips and nose. Such movements are often related to the expression of specific emotions. FAUs have been already used to train neural networks and they are the key features of the top-scoring baseline method released by the MuSe challenge organizers. For these reasons, we leverage the FAUs extracted from N selected frames to enhance the visual component representation.

Similar to the baseline method, we use *Py-Feat*[4] to extract 20 different FAUs from the selected frames, but we exploit the logistic regressor pretrained model instead of the random forest because deprecated. Moreover, the tool provides the scores of 7 emotions (i.e., anger, disgust, fear, happiness, sadness, surprise and neutral), which are slightly different from those proposed by the task. Hence, we include the aforesaid scores in the FAU feature set.

3.5 Domain knowledge for recognizing emotions

Front-facing camera videos, commonly a webcam recording, characterize the MuSe-Reaction dataset. This is a peculiar characteristics of the input data, which led us to make the following considerations on the extracted visual features:

- To recognize emotions we primarily target the facial expression.
- Since the video recordings are self-made by the human subject in an uncontrolled environment, the background likely includes noisy elements that might divert the model's attention from the facial expression.
- We need a way to crop faces from the video frames.

To our aim, we use YOLO5Face [22] to automatically detect faces within the frames and crop them to remove background information. Moreover, to make our feature extractor more capable of handling face images, we use transfer learning to inject some prior knowledge about physiognomy into the model. In addition to the base model pretrained on ImageNet [9], we include in our experiments the feature extracted by a ViT [1] pretrained on an age estimation task using FairFace [14] (i.e., ViT-Face), a close-up photo dataset.

3.6 Attention-based late fusion

We try several combinations of feature sets, including image-related ones, to solve the downstream task. For each combination, we concatenate the involved sets of features to form a unique input array for each of the N selected frames per video. In such a way, we obtain a sequence of N input tokens of fixed length, depending on the features involved. Then, we use these arrays to feed a Perceiver[13] model to perform the multi-regression task.

Perceiver is an attention-based modality-agnostic model that can merge information from different input modalities without making any modality-specific assumption. In contrast with other late-fusion approaches, the attention bottleneck mechanism in the Perceiver architecture ensures its scalability, i.e., it transforms the inputs into a fixed-length latent representation that can be processed with a computational cost unaffected by the original input size.

We set the number of neurons in the last linear layer equal to the number of task-related emotions. Thus, a score can be assigned to each of them with a single instance of our architecture.

4 EXPERIMENTAL SETTINGS

The challenge organizers provide access to the Hume-Reaction dataset containing fine-grained annotations for 75 hours of video clips showing emotional reactions. Each video is self-annotated with seven scores, one for each emotion, normalized to the $[0, 1]$ range. The data collection comprises 25067 audio-visual clips, each one lasting from 10 to 15 seconds and with an average duration of 11.63 seconds.

A video is a sequence of frames reproduced at a Frame Per Second (FPS) of 30, which gives the viewer the illusion of fluid movement. The overall number of frames per video varies from 300 to 450. For this reason, in all our experiments, we chose 32 as the number of selected frames N. We come up with, on average, one frame every 0.37 seconds of video and an average audio fragment length of 0.74 seconds.

The Perceiver for late fusion is fine-tuned using the average Mean Squared Error (MSE) loss among emotions. The late fusion module is trained for 50 epochs, using a batch size of 16 with an initial learning rate of 10^{-5} halved every 10 epochs. We select the best checkpoint according to the Pearson correlation score on the development set.

Modality contributions. To investigate the contribution of each modality to the overall performance, we train and evaluate ViPER in different settings.

- *Visual-only*: ViPER is trained using only the image embeddings extracted by ViT.
- *Bi-modal*: ViPER is trained using the image embeddings combined with one of the other modalities (i.e., audio, text, or FAU), independently.
- *Tri-modal*: ViPER is trained using the image embeddings combined with two additional modalities.
- *Full*: ViPER is trained using image, audio and text encodings combined with Facial Action Units.

Hardware settings. The experiments were performed on a cluster equipped with Intel® Xeon® Scalable Processor Gold 6130 dual-CPU, Nvidia® Tesla® V100 GPU, and 384 GB of shared RAM, running CentOS 7.6.

5 RESULTS

In this section we present and discuss the results of ViPER in the MuSe-Reaction Challenge. Specifically, Section 5.1 reports the in-depth analysis of the impact of the visual component, whereas

[3] https://anonymous.4open.science/r/MuSe2022_reaction-57E7
[4] https://py-feat.org Latest access: July 2022

Table 2: Performance comparison between the ViPER settings and the baseline methods relying on the visual encoders only. Performance are reported in terms of Pearson correlation.

Model	Encoder	Face crop	Dev set
FAU [8]	-	-	**0.2840**
VGGFace2 [8]	-	-	0.2488
ViPER	ViT-Base	-	0.2102
	ViT-Face	-	0.2223
	ViT-Base	✓	0.2580
	ViT-Face	✓	0.2712

Section 5.2 discusses the outcomes of an ablation study aimed at exploring the impact of each modality on the emotion prediction task. Finally, Section 5.3 discusses the quality of the per-emotion prediction outcomes.

5.1 Impact of the visual features

Visual encoders provide a high-level description of the dependencies among video frame contents. Table 2 compares the performance of (1) organizers' provided baselines (i.e., FAU and VGGFace2 [8]), (2) different ViPER settings based on various pre-trained ViT models, both including face cropping strategies and not, (3) a standard pre-trained model (i.e., ViT) with ViT-Face, which is a specific version of ViT fine-tuned on FairFace dataset [14] for age estimation.

Using a pre-trained ViT model fine-tuned on face images (ViT-Face) improves the performance of ViPER compared to that of the standard ViT pre-trained model. According to these findings, using a model pre-trained on face images allows us to extract features that better represent human emotions, which in turn allows us to achieve more accurate emotion predictions.

The results confirm that by cropping the face area within each video frame and by forcing the learning process to focus solely on the facial expressions, ViPER significantly improves its performance compared to an entire video frame analysis. This also implies that the features extracted by the visual component are not able to sufficiently represent emotional expressions if the model is allowed to process the whole video frame, which is likely due to the presence of background information that might distract the model from focusing on facial expressions. Therefore, using the induction bias provided by the face cropping strategy might help the model extract features that better represent emotional expressions.

5.2 Ablation study on the MuSe-Reaction challenge

Table 3 summarizes all the ViPER results achieved with all the tested settings as well as the outcomes of the baseline methods released by the challenge organizers (i.e., FAU and VGGFace2 [8]).

All the system settings were tested in the development set. Challenge participants were allowed to submit only 5 settings for the evaluation on the test set. To gain insights into both separate and joint modality contributions, we submitted the *visual-only* and *bi-modal* systems to the evaluation platform. The results achieved on the test set are reported in Table 3.

From the comparison between single-modal and bi-modal system performance, we can observe that bi-modal settings yield significant improvements. Specifically, their combination of the image embeddings computed using the Facial Action Units performs best among the bi-modal settings. Audio embeddings turn out to be the least predictive features both in the development and the test set.

Because of the enforcement of the maximum number of submissions per team (5), we miss the outcomes of the tri-modal setting on the test set. However, the results on the development set, clearly indicate a substantial improvement compared to the bi-modal models. Similar to the former cases, the integration of image embeddings and FAU is highly beneficial.

Text embeddings are less discriminative than image embeddings but also more important than acoustic ones. The main reason is that video recordings do not contain enough acoustic features to be properly exploited by the automatic audio processing. In fact, human subjects silently watch the video and never talk throughout the video stream, thus limiting the amount of information that can be extracted from the audio. Moreover, the WavLM model used for audio embeddings has been trained on speech, thus it may not be effective for encoding audio signals in our scenario. However, to better understand the contribution of text-based features, we plan to experiment with different architectures and data representations for the text modality.

Finally, the full configuration including image, audio, text embeddings and Facial Action Units achieves the best performance among all the other settings, both on the development set and on the test set. This demonstrates the effectiveness of the Perceiver-based late fusion approach to the emotion prediction task.

5.3 Analysis of emotion-specific performance

The prediction outcomes show a relevant variability across emotions. This indicates that ViPER is able to better estimate the scores associated to some specific emotion classes (e.g., Amusement, Anxiety, Fear) whereas struggles to classify other ones (e.g., Adoration, Disgust, Empathic-Pain, Surprise).

Figures 2 and 3 show the confusion matrices for the best and worst performing emotion classes (i.e., Amusement and Empathic-Pain). For the Amusement class, most of the matrix values fall in the $[0.4, 1]$ range, whereas the predictions for the Empathic-Pain class mainly fall into the $[0.1, 0.4]$ range. This could be due to the fact that the model is more effective in providing a fine-grained estimates for the intensity of the *Amusement* class, whereas is less accurate to predict the *Empathic-Pain* class. In a nutshell, extensive prediction ranges likely imply higher prediction accuracy, whereas smaller prediction ranges are a strong clue of the difficulties in emotion intensity estimation. While considering all the emotion classes together, the average absolute value of the predictions is uncorrelated with the overall emotion class performance.

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we proposed ViPER, a multimodal approach to predict the emotional state of a user reacting to a video. The proposed approach relies on a modality-agnostic late fusion network, which provides ViPER users with a flexible and easy-to-adapt framework.

Table 3: Performance comparison of ViPER model in multimodal settings. Performance are reported in terms of Pearson correlation. Best performing ViPER setting is statistically significant (p-value < 0.01) with respect other ViPER configurations.

Model	Image	Text	Audio	FAU	Dev set	Test set
FAU [8]	-	-	-	✓	0.2840	0.2801
VGGFace2 [8]	✓	-	-	-	0.2488	0.1830
ViPER	✓	-	-	-	0.2712	0.2622
	✓	✓	-	-	0.2806	0.2702
	✓	-	✓	-	0.2748	0.2610
	✓	-	-	✓	0.2978	0.2859
	✓	✓	✓	-	0.2854	-
	✓	✓	-	✓	0.3014	-
	✓	-	✓	✓	0.2924	-
	✓	✓	✓	✓	**0.3025**	**0.2970**

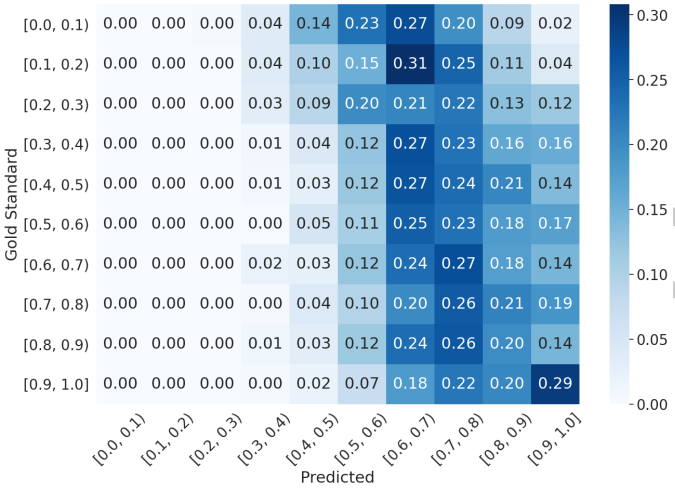

Figure 2: Amusement prediction analysis. $\rho = 0.3414$

The effectiveness of ViPER has been demonstrated on the Hume-Reaction dataset proposed for the MuSe 2022 Reaction sub-challenge. The empirical findings indicate that all the examined modalities positively contribute to the overall performance of the proposed approach. In addition, the integration of domain knowledge, i.e., face cropping and in-domain ViT pre-training, turns out to be beneficial as well.

As future work, we plan to extend our approach by adding a pre-training step, in which modality-specific encoders are trained using in-domain data collections to enable a more discriminating feature extraction step for the emotion recognition task. Since the integration of ViT-Face has shown to be particularly effective among the visual features, we also plan to test this setting for the other input modalities. The benefits of using template-based frame captioning techniques to augment data with textual annotations produced

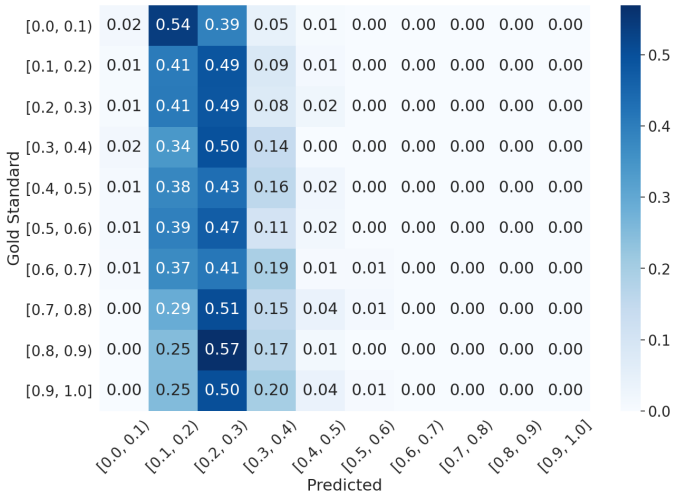

Figure 3: Empathic-Pain prediction analysis. $\rho = 0.2547$

satisfactory results. Hence, we will also explore multiple ways to further extend the video descriptions. This is also instrumental for the modality-agnostic fusion network, which is inherently capable of integrating additional annotations.

Finally, we aim at comparing the proposed methodology with other state-of-the-art fusion approaches, as well as explore other fusion strategies. For the latter, our goal is to investigate modality fusion at different levels of the architecture. Since the proposed late fusion strategy can be suboptimal for specific input modalities, we plan to leverage a recently proposed technique to better estimate the fusion point in the architecture [17].

ACKNOWLEDGMENTS

The research leading to these results has been partly funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino[5].

REFERENCES

[1] 2022. nateraw/vit-age-classifier · Hugging Face. https://huggingface.co/nateraw/vit-age-classifier [Online; accessed 1. Jul. 2022].
[2] MAH Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. 2021. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* 10, 9 (2021), 1036.
[3] Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 Task 5: Using Perceiver IO for Detecting Misogynous Memes with Text and Image Modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.
[4] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards real-time speech emotion recognition for affective e-learning. *Education and information technologies* 21, 5 (2016), 1367–1386.
[5] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134.
[6] Alessandro Bondielli and Lucia C Passaro. 2021. Leveraging CLIP for Image Emotion Recognition.. In *NL4AI@ AI* IA*.

[5] http://www.hpc.polito.it

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900* (2021).

[8] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Mueller, Lukas Stappen, Eva Messner, Andreas König, Alan Cowen, Erik Cambria, and Björn Schuller. 2022. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. (04 2022).

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[11] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.

[12] Jad Haddad, Olivier Lézoray, and Philippe Hamel. 2020. 3d-cnn for facial emotion recognition in videos. In *International symposium on visual computing*. Springer, 298–309.

[13] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*. PMLR, 4651–4664.

[14] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[16] Fuyan Ma, Bin Sun, and Shutao Li. 2021. Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing* (2021).

[17] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are Multimodal Transformers Robust to Missing Modality?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18177–18186.

[18] Ninad Mehendale. 2020. Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences* 2, 3 (2020), 1–8.

[19] Asif Iqbal Middya, Baibhav Nag, and Sarbani Roy. 2022. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems* 244 (2022), 108580.

[20] Trinh Thi Doan Pham and Chee Sun Won. 2019. Facial Action Units for Training Convolutional Neural Networks. *IEEE Access* 7 (2019), 77816–77824. https://doi.org/10.1109/ACCESS.2019.2921241

[21] Francisco A Pujol, Higinio Mora, and Ana Martínez. 2019. Emotion recognition to improve e-healthcare systems in smart cities. In *The International Research & Innovation Forum*. Springer, 245–254.

[22] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. 2021. YOLO5Face: Why Reinventing a Face Detector. (2021).

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[24] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. 2013. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology* 16, 2 (2013), 143–160.

[25] Muhammad Sajjad, Sana Zahir, Amin Ullah, Zahid Akhtar, and Khan Muhammad. 2020. Human behavior understanding in big multimedia data using CNN based facial expression recognition. *Mobile networks and applications* 25, 4 (2020), 1611–1621.

[26] Björn W Schuller, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, Sebastian P Bayerl, Korbinian Riedhammer, Adria Mallol-Ragolta, et al. 2022. The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitoes. *arXiv preprint arXiv:2205.06799* (2022).

[27] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5329–5333.

[28] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-Modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop* (Seattle, WA, USA) *(MuSe'20)*. Association for Computing Machinery, New York, NY, USA, 27–34. https://doi.org/10.1145/3423327.3423672

[29] Licai Sun, Mingyu Xu, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. 2021. Multimodal Emotion Recognition and Sentiment Analysis via Attention Enhanced Recurrent Model. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge* (Virtual Event, China) *(MuSe '21)*. Association for Computing Machinery, New York, NY, USA, 15–20. https://doi.org/10.1145/3475957.3484456

[30] Ivona Tautkute, Tomasz Trzcinski, and Adam Bielski. 2018. I Know How You Feel: Emotion Recognition With Facial Landmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[31] Shao-Yen Tseng, Shrikanth Narayanan, and Panayiotis Georgiou. 2021. Multimodal embeddings from language models for emotion recognition in the wild. *IEEE Signal Processing Letters* 28 (2021), 608–612.

[32] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing* 11, 8 (2017), 1301–1309.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Matthias Wimmer, Björn Schuller, Dejan Arsic, Bernd Radig, and Gerhard Rigoll. 2008. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proc. 3rd Int. Conf. on Computer Vision Theory and Applications VISAPP, Funchal, Madeira, Portugal*. 145–151.

[35] Karsten Wolf. 2015. Measuring facial expression of emotion. *Dialogues in Clinical Neuroscience* 17, 4 (2015), 457–462. https://doi.org/10.31887/DCNS.2015.17.4/kwolf

[36] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-Perceiver: Pre-Training Unified Architecture for Generic Perception for Zero-Shot and Few-Shot Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16804–16815.