PainFusion: Multimodal Pain Assessment from RGB and Sensor Data

Benavent-Lledo,M., Lopez-Valle, M., Ortiz-Perez, D., Mulero-Perez, D., Garcia-Rodriguez, J. and Psarrou, A.

PainFusion: Multimodal Pain Assessment from RGB and Sensor Data

Manuel Benavent-Lledo[1], Maria Dolores Lopez-Valle[1], David Ortiz-Perez[1], David Mulero-Perez[1], Jose Garcia-Rodriguez[1], and Alexandra Psarrou[2]

[1] Department of Computer Technology, University of Alicante, Alicante, Spain
{mbenavent,mdlopez,dortiz,dmulero,jgarcia}@dtic.ua.es
[2] School of Computer Science and Engineering, University of Westminster, London, UK
psarroa@westminster.ac.uk

Abstract. Traditional pain assessment tools often rely on subjective self-reporting methods, hindering the work of healthcare professionals. However, the patient's facial expressions and biomedical data provide a reliable source of information for caregivers. In this work, we present a multimodal architecture that utilizes both RGB video and biomedical sensor data from the BioVid Heat Pain dataset. We use video transformer architectures in conjunction with a thorough analysis of biomedical signals, including galvanic skin response, electromyography, and electrocardiogram, for comprehensive feature extraction. These features are then fused to create a robust model for pain assessment. Experimental results show that our multimodal architecture outperforms unimodal video-based methods in pain detection. Furthermore, our study highlights the potential of combining non-invasive video analysis with physiological data to facilitate pain prediction and management in clinical settings, paving the way for more accurate and efficient pain assessment methods that can be used in various healthcare applications.

Keywords: pain assessment; computer vision; deep learning; sensor data; signal processing

1 Introduction

Pain, an inevitable part of human experience, arises from injury, illness, or medical interventions, often prompting individuals to seek primary care [4]. While most pain does not escalate to chronic levels, its subjective nature and variability—shaped by factors such as gender, age, religious beliefs, and ethnic background—render it a significant concern in healthcare [26]. Chronic pain, in particular, profoundly affects daily functioning and emotional well-being, underscoring the need for comprehensive management to improve patients' quality of life. Accurate pain prediction is critical, especially for those who struggle to communicate their pain effectively.

Predicting pain not only aids in its management but also enhances medical care through early and personalized interventions. This approach optimizes

therapeutic strategies and supports multidisciplinary efforts to address the complexities of pain. Current tools for assessing pain, such as pain scales and questionnaires [12,23], offer standardized methods to quantify pain intensity and its impact on life. However, these tools rely heavily on patient self-reporting, which introduces biases and variability, limiting their predictive accuracy.

Physiological measures present a promising avenue for enhancing pain prediction. Healthcare professionals and researchers can obtain a comprehensive view by understanding the underlying mechanisms of pain through the integration of various technologies. Tools such as functional neuroimaging, biosensors, and electromyography (EMG) offer objective data on pain-related physiological changes [18]. For example, functional neuroimaging reveals the neuronal circuits involved in pain perception, while biosensors measure physiological indicators like heart rate and skin conductance. EMG provides insights into muscle activity and nerve function, crucial for understanding pain's impact on the body.

Recognizing the limitations of traditional methods and the potential of physiological approaches, this study employs the BioVid Heat Pain Database [34] to develop a multimodal model based on a transformer architecture. By examining correlations from biomedical sensors and utilizing computer vision techniques for facial expression analysis, we aim to advance pain assessment methodologies. As a result, we propose the multimodal architecture PainFusion for pain assessment on 5 different levels, 0 to 4, where 0 represents no pain and 4 represents the most severe pain.

The remaining of this paper is organized as follows. Section 2 summarizes relevant work for video understanding and pain assessment. Section 3 presents the proposed architecture, PainFusion. Experiments and results are detailed in Section 4. Finally, conclusions from this work are drawn in Section 5.

2 Related Work

This section provides an overview of current methodologies for addressing the pain estimation problem. Previous work is broad with diverse approaches such as [33,24,16,10,25,21]. Particularly, pain assessment modalities can be divided into two main categories: behavioral and physiological. Behavioral modalities include facial expressions, body movements (such as guarding, rubbing, restlessness, and head movements), vocalizations (like crying or moaning), and spoken words, which can be transcribed via speech recognition to capture self-reported information. Physiological modalities encompass brain activity, cardiovascular activity, and electrodermal activity.

Among the existing methods, unimodal approaches consist of leveraging a single modality as input. One of the leading methodologies to address this challenge comprises video footage from the face of the person. In this area, convolutional neural networks (CNNs) are widely adopted and extensively discussed in [28]. Methods such as SANET [9] and SDNET [19] are employed, with SANET excelling in automatically identifying spatial attributes like color, and SDNET specializing in extracting shape-related features such as facial contours. These

networks are crucial in pre-processing input images and extracting significant features, which are then processed in a learning phase to compute a pain score according to the Prkachin and Solomon Pain Intensity (PSPI) scale [23].

The authors in [6] explore the use of the self-attention mechanism from transformer architectures. Firstly introduced for natural language processing [31], vision transformers [8,5] have demonstrated remarkable capabilities for feature extraction. However, as remarked by the authors, analyzing a single frame lacks temporal context provided by video inputs. Due to this fact, video transformer architectures [2,29,7,27] were used during experiments on the BioVid dataset [34] and are also explored, providing better results than single-frame analysis.

Alternatively, instead of RGB data, the use of biomedical sensors to quantitatively measure physiological signals is an extended approach, albeit more invasive to the user than the previous one. Typically, signals are fused in multimodal approaches as presented in prior work. For example, in [35,36] facial expressions are combined with head poses. Alternatively, [36,1] fuse EDA, ECG and sEMG signals, and [36,13] leverage the same information with the addition of video inputs. Kessler et al. [14] propose an architecture that uses video, RSP, ECG and remote PPG. Audiovisual inputs are especially relevant if we are able to analyze how a person's speech changes when subjected to pain [30]. Or body movements from motion capture and sEMG as the study conducted in [20].

The study on pain, its types, and traditional prediction methodologies emphasizes the importance of selecting appropriate datasets for applying machine learning techniques. Among the most relevant ones we find the Delaware Pain Database [17], which contains photographs of pain expressions; BP4D [15], featuring a wide range of multimodal data on spontaneous emotions; MIntPAIN [11], focusing on pain levels via visual, depth, and thermal information; EmoPAIN [3], which captures natural facial expressions and body movements of chronic pain sufferers; Sense Emotion [32], distinguishing pain from emotions through multimodal sensory data; X-ITE [22], evaluating pain intensity with various sensor recordings; and BioVid Heat Pain Database [34], using heat stimuli to study pain responses. Each dataset offers unique attributes crucial for building effective predictive models.

3 PainFusion Architecture

In this section we present our PainFusion architecture (Fig. 1) for pain assessment from multimodal data. The following subsections detail each of the components of the PainFusion architecture designed for pain assessment on the BioVid dataset [34].

3.1 BioVid Heat Pain Database

The BioVid dataset [34] contains videos of 90 healthy adults between the ages of 20 and 65. During controlled experiments, participants were exposed to thermal stimuli on different body regions, such as the forearm and leg. The dataset is

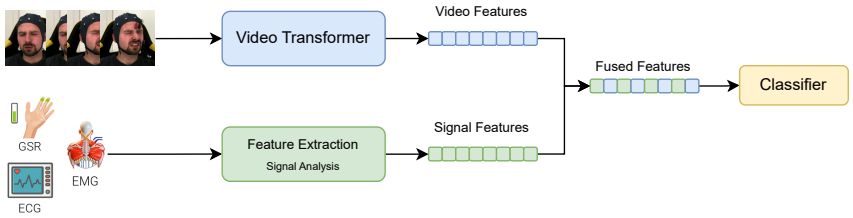

Fig. 1. Overview of the PainFusion. We first extract video features from RGB video clips from the face of the patient using a video transformer architecture. These features are then fused with pre-extracted features from biomedical signals comprising GSR, ECG and EMG. The resulting vector is fed onto a classifier for discrimination between no pain and 4 different pain levels.

remarkably diverse, providing detailed participant information, including age, gender, medical history, and pain sensitivity. It also documents subjective pain responses, such as ratings of pain intensity and discomfort, as well as objective data such as ECG, EMG, skin conductance, and heart rate.

The dataset is divided into five subsets, each containing different types of data. This study focuses on subset *A*, which provides a balanced collection of data across five classes: BL0 (no pain) and PA1 to PA4, representing increasing levels of pain experienced by the participants.

3.2 Video Transformer

As demonstrated by prior work [6,27], video transformers hold remarkable capabilities for feature extraction given their ability to model temporal and spatial information from video frames adequately. In this work, as in [6], we exploit 3 well-known video transformer architectures: TimeSformer, ViVit and VideoMAE.

TimeSformer *[7].* This architecture, based on the original transformer architecture introduced for natural language processing [31], is specifically designed for handling temporal sequences. This model adapts the Vision Transformer [8] principles for video processing by extending them into the temporal domain. It relies entirely on self-attention layers, eliminating the need for convolutional layers. A comparative analysis of different attention mechanisms within this model was performed by the authors. The *divided attention* approach, which uses separate temporal and spatial attention in different network blocks, yielded the most favorable results and is the one used in our experiments.

ViVit *[2].* The authors of this architecture presented a novel approach in response to the success of transformers in the image domain. Four different models were proposed depending on how temporal modeling is performed for the video classification task:

- *Transformer-Encoder:* this architecture comprises an extension of the Vision Transformer (ViT) [8] into the temporal domain.
- *Factorized Encoder:* unlike the naive transformer-encoder, this approach does not use a single encoder for all videos. Instead, each video is divided into multiple chunks. Spatial attention is first applied to each clip, and the resulting vectors are fed into a second encoder with a temporal focus. Positional encoding is used for proper identification of video clips, *i.e.* an index is assigned for identification.
- *Factorized Self-Attention:* this approach presents a similar architecture to the Transformer-Encoder although introducing a two-stage attention computation. First spatial attention for frame level video features, and second temporal attention for modeling of past events.
- *Factorized Dot-Product:* the naive version of the Transformer-Encoder is leveraged in this approach after low-level modifications. In the attention layers, the dot-product operations are divided so that half of the operations are performed on spatial tokens and the remaining half on temporal ones.

Among these architectures, the Factorized Encoder yields the best results and consequently is the one used for experimentation in this work.

VideoMAE *[29].* While not a transformer itself, this method leverages transformer based models to achieve state-of-the-art results. In their paper authors explore the use of autoencoders for the self-supervised pre-training task essential for transformers. Frame-level features are extracted using the aforementioned ViT [8].

3.3 Biomedical Signal Analysis

We conduct an extensive analysis over the biomedical signal data provided in the BioVid dataset [34], extending the analysis over this signals conducted in [6], we extract the features from each signal as follows:

Galvanic Skin Response (GSR). For the analysis of GSR signals the following features are extracted:

- **Maximum Value ($G_{\mathbf{max}}$)**: The highest value within the signal.
- **Mean Value ($\bar{G}$)**: The average value of the signal, calculated as:

$$\bar{G} = \frac{1}{N} \sum_{i=1}^{N} G_i$$

where G_i represents the individual signal values and N is the total number of samples.

– **Slope Changes**: These are detected using the second derivative of the signal. Changes in the sign of $\frac{d^2G}{dt^2}$ indicate alterations in the slope. To formally detect these changes, we define the second derivative at sample i as:

$$\left.\frac{d^2G}{dt^2}\right|_i \approx G_{i+1} - 2G_i + G_{i-1}$$

Slope changes occur where the sign of $\frac{d^2G}{dt^2}$ changes:

$$\text{Slope Change at } i \text{ if } \left(\left.\frac{d^2G}{dt^2}\right|_i \cdot \left.\frac{d^2G}{dt^2}\right|_{i-1} < 0\right)$$

Smoothing is applied to the signal to reduce noise and avoid detecting excessive slope changes.

Electromyogram (EMG). The EMG signal analysis included the following features:

– **Mean ($\bar{E}$)**: The average value of the EMG signal.
– **Variance (σ_E^2)**: The measure of signal variability, given by:

$$\sigma_E^2 = \frac{1}{N}\sum_{i=1}^{N}(E_i - \bar{E})^2$$

– **Kurtosis**: A statistical measure of the "tailedness" of the signal distribution. In other words, we measure the shape of the signal's distribution, particularly focusing on the presence and extremity of outliers in the data.
– **Peak Detection**: Peaks were identified using a threshold set at half the maximum amplitude:

$$E_{\text{threshold}} = 0.5 \cdot E_{\max}$$

– **Slope Changes**: Identified using the first derivative $\frac{dE}{dt}$. Slope changes occur where the sign changes:

$$\text{Slope Change at } i \text{ if } \left(\left.\frac{dE}{dt}\right|_i \cdot \left.\frac{dE}{dt}\right|_{i-1} < 0\right)$$

– **Peak Density**: The number of peaks per second, considering a sampling frequency of 512 Hz.

Electrocardiogram (ECG). For the last signal the extracted features are as follows:

– **Mean Value ($\bar{C}$)**: The average value of the ECG signal.
– **Peak Detection**: Key peaks (R, P, and T) were identified after slight smoothing of the signal. The R peak corresponds to the depolarization of the ventricles, indicating ventricular contraction. The P peak represents atrial depolarization, occurring just before atrial contraction. Lastly, the T peak signifies ventricular repolarization, occurring as the ventricles prepare for relaxation after contraction.

– **Intervals and Amplitudes**: Distances between peaks (*e.g.* RR interval) and their amplitudes were calculated. The mean values for both intervals and amplitudes were computed. The heart rate was derived from the RR interval.

The resulting vector contains 27 features that represent the lower branch in Figure 1 (colored in green).

3.4 Training

As depicted in Figure 1, after extracting the features, these are fused via concatenation resulting in a single vector that may be used for classification. To this end, we employ a Multi-Layer Perceptron (MLP) classifier with Rectified Linear Unit (ReLU) activation and dropout to avoid overfitting. We use cross-entropy loss to measure the classification performance.

Formally, let $\mathbf{z}$ be the feature vector fed to the classifier whose parameters are denoted by θ. The output of the MLP, $\mathbf{y} = f(\mathbf{z}; \theta)$, represents the predicted class probabilities. The cross-entropy loss $\mathcal{L}$ is used to evaluate the discrepancy between the predicted class probabilities and the true class labels. Let $\mathbf{y}_{\text{true}}$ be the one-hot encoded true class labels. The cross-entropy loss is given by:

$$\mathcal{L}(\mathbf{y}, \mathbf{y}_{\text{true}}) = -\sum_{i=1}^{C} y_{\text{true},i} \log(y_i)$$

where C is the number of classes, $y_{\text{true},i}$ is the true label for class i, and y_i is the predicted probability for class i.

4 Experiments

This section details the conducted experiments using the BioVid dataset [34], training details and presents the results obtained from the various modalities.

4.1 Experimental setup

During experiments we split the BioVid dataset [34] in 3 non-overlapping splits for training (80%), validation (10% of training data) and testing (remaining 20% of data). We present the results over the validation and test splits. Accuracy is used as evaluation metric to measure the performance of the models.

The proposed method has been implemented in PyTorch and all experiments are performed on a NVIDIA RTX 3090 GPU. As a result, there is a limitation when training video models and we are able to use a maximum batch size of 4. Notably, when using signal features only the batch size is increased to 16. We train the models for 20 epochs with a learning rate of 5×10^{-7} and AdamW optimizer with a decay of 0.01. Video frames undergo a preprocessing step consisting of a downsampling and defining input windows of 32, 16 and 8 to leverage pretrained models from Vivit [2], VideoMAE [29] and Timesformer [7], respectively.

Our implementation has been open-sourced and may be found in our GitHub repository[3].

4.2 Results

Table 1 presents the results obtained from the experiments. We can conclude that the features extracted from biomedical signals provide the best result with a classification accuracy of 62.21%. Additionally, in the test subset of Timesformer and the validation subset of Vivit and VideoMAE, the use of biomedical signals improve the results of the unimodal video input.

Video Transformer	Biomedical Signals	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
-	✓	0.60723	0.21624	**0.62213**	0.16954
TimeSformer		0.56598	0.21121	0.43467	0.20632
	✓	0.48789	0.20690	0.46506	0.21437
Vivit		0.43500	0.22557	0.40485	0.22471
	✓	0.50606	0.21264	0.38709	0.22759
VideoMAE		0.48192	0.22629	**0.50686**	0.21724
	✓	0.55433	0.22917	0.48059	0.53793

Table 1. Performance comparison of different models on Validation and Test sets.

Despite the lower results for video approaches, it is worth noting the advantages of video approaches over biosensors. While heart rate information can be obtained from simple smartbands on the patient's wrist, EMG and GSR sensors require the participant to be connected to the device. Video approaches, on the other hand, are less invasive, requiring only a camera pointed at the patient's face.

5 Conclusions

This paper presents a multimodal architecture for pain assessment from RGB videos and biomedical sensor data including galvanic skin response, electromyogram, and electrocardiogram. After a comprehensive review of state-of-the-art methods, we present PainFusion, an architecture that fuses video transformer features with manually extracted features after extensive analysis on the provided signals. The results show that the multimodal architecture improves the results compared to the unimodal video approach. However, these results present a lower accuracy compared to the unimodal signal approach. Despite the results obtained, the use of video as a modality is emphasized for its non-invasive nature and speed in pain detection, reducing the risks associated with invasive procedures and facilitating faster evaluations in medical care.

[3] https://github.com/3dperceptionlab/tfg_mdlopez

Future work in this area may explore binary approaches for prediction as prior work to detect the main objective of this problem, gathering information whether the person is in pain or not. Similarly, exploring other datasets with different signals such as EEG may be of interest. Finally, it is worth mentioning that data analysis can be costly thereby, future work may also explore the use of algorithms for automatic feature extraction (*e.g.* using convolutions on signal data) or experimenting with other fusion strategies, such as cross-modal fusion or early fusion apart from the late fusion strategy proposed in this work.

Acknowledgment

We would like to thank CIAICO/2022/132 Consolidated group project "AI4-Health" funded by the Valencian government. This work has also been supported by a Spanish national and a regional grants for PhD studies, FPU21/00414, CIACIF/2021/430 and CIACIF/2022/175.

References

1. Amirian, M., et al.: Using Radial Basis Function Neural Networks for Continuous and Discrete Pain Estimation from Bio-physiological Signals, p. 269–284. Springer (2016)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proc. IEEE/CVF ICCV, pp. 6836–6846 (2021)
3. Aung, M.S.H., et al.: The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal emopain dataset. IEEE Transactions on Affective Computing **7**(4), 435–451 (2016)
4. Babarro, A.A.: La importancia de evaluar adecuadamente el dolor. Atención primaria **43**(11), 575 (2011)
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv:2106.08254 (2021)
6. Benavent-Lledo, M., et al.: A comprehensive study on pain assessment from multimodal sensor data. Sensors **23**(24) (2023)
7. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
9. Fan, H., Ling, H.: Sanet: Structure-aware network for visual tracking (2017)
10. Gomez-Donoso, F., et al.: A robotic platform for customized and interactive rehabilitation of persons with disabilities. Pattern Recognit. Lett. **99**, 105–113 (2017)
11. Haque, M.A., et al.: Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities. In: FG, pp. 250–257 (2018)
12. Ibáñez, R.M., et al.: Escalas de valoración del dolor. Jano **25**(1), 41–4 (2005)
13. Kächele, M., et al.: Multimodal data fusion for person-independent, continuous estimation of pain intensity. In: Eng. App. of Neural Networks, pp. 275–285. Springer (2015)
14. Kessler, V., Thiam, P., Amirian, M., Schwenker, F.: Pain recognition with camera photoplethysmography. In: IPTA. IEEE (2017)

15. Li, X., Zhang, X., Yang, H., Duan, W., Dai, W., Yin, L.: An eeg-based multimodal emotion database with both posed and authentic facial actions for emotion analysis. In: Face and Gestures, pp. 336–343 (2020)
16. López, J.A., et al.: A novel prediction method for early recognition of global human behaviour in image sequences. Neural Process. Lett. **43**(2), 363–387 (2016)
17. Mende-Siedlecki, P., et al.: The delaware pain database: a set of painful expressions and corresponding norming data. PAIN Reports **5**(6), e853 (2020)
18. Moreno-Serrano, N.L.R., et al.: Medicina del dolor y cuidado paliativo. Editorial Universidad del Rosario (2022)
19. Ochs, M., Kretz, A., Mester, R.: Sdnet: Semantically guided depth estimation network (2019)
20. Olugbade, T.A., et al.: Bi-modal detection of painful reaching for chronic pain rehabilitation systems. In: Int. Conf. on Multimodal Interaction (2014)
21. Ortiz-Perez, D., Ruiz-Ponce, P., Tomás, D., Garcia-Rodriguez, J., Vizcaya-Moreno, M.F., Leo, M.: A deep learning-based multimodal architecture to predict signs of dementia. Neurocomputing **548**, 126,413 (2023)
22. Othman, E., et al.: Automatic vs. human recognition of pain intensity from facial expression on the x-ite pain database. Sensors **21**(9), 3273 (2021)
23. Prkachin, K.M., Solomon, P.E.: The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. Pain **139**(2), 267–274 (2008)
24. Revuelta, F.F., et al.: Representation of 2d objects with a topology preserving network. In: 2nd Int. Workshop on Pattern Recognition in Information Systems, April 2002, pp. 267–276 (2002)
25. Ruiz-Ponce, P., et al.: POSEIDON: A data augmentation tool for small object detection datasets in maritime environments. Sensors **23**(7), 3691 (2023)
26. Santiago, A.J., Sánchez, S.B.: Experiencia diferencial del dolor según género, edad, adscripción religiosa y pertenencia étnica. Archivos en Medicina Familiar **16**(3), 49–55 (2017)
27. Selva, J., et al.: Video transformers: A survey. TPAMI (2023)
28. Semwal, A., et al.: Computer aided pain detection and intensity estimation using compact CNN based fusion network. Applied Soft Computing **112**, 107,780 (2021)
29. Tong, Z., et al.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS **35**, 10,078–10,093 (2022)
30. Tsai, F.S., Hsu, Y.L., Chen, W.C., Weng, Y.M., Ng, C.J., Lee, C.C.: Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions. In: Interspeech 2016. ISCA (2016)
31. Vaswani, A., et al.: Attention is all you need. NeurIPS **30** (2017)
32. Velana, M., et al.: The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In: MPRSS Workshop, pp. 127–139 (2017)
33. Viejo, D., et al.: Using GNG to improve 3d feature extraction - application to 6dof egomotion. Neural Networks **32**, 138–146 (2012)
34. Walter, S., et al.: The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE International Conference on Cybernetics (CYBCO), pp. 128–131 (2013)
35. Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., Traue, H.C.: Automatic pain assessment with facial activity descriptors. IEEE Transactions on Affective Computing **8**(3), 286–299 (2017)
36. Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., Traue, H.C.: Automatic pain recognition from video and biomedical signals. In: ICPR 2014 (2014)