MMGT: MULTIMODAL GRAPH-BASED TRANSFORMER FOR PAIN DETECTION

Kevin Feghoul [1,2]*, Deise Santana Maia* [2]*, Mohamed Daoudi* [2,3]*, Ali Amad* [1]

[1]Univ. Lille, Inserm, CHU Lille, UMR-S1172 LilNCog, F-59000 Lille, France
[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3]IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France
{kevin.feghoul, deise.santanamaia, mohamed.daoudi, ali.amad}@univ-lille.fr

ABSTRACT

Pain can be expressed from multiple modalities, such as facial expressions, physiological signals, and behaviors. For that reason, multimodal learning can greatly benefit automatic pain detection and, more generally, a variety of tasks in the field of affective computing. In this context, as one of our main contributions, we leverage the multimodal interaction among the intermediate modality representations, which are rarely exploited in existing works. In order to capture the relationships between multiple modalities, we propose the Multimodal Graph-based Transformer (MMGT), in which unimodality feature extraction is performed using Transformers and then fused using a Graph Convolutional Network (GCN). We evaluated MMGT on the BP4D+ dataset, and the results demonstrate the efficiency of our fusion framework for the task of pain detection, which outperformed all the existing approaches under multimodal settings. Our best results were obtained using 2D facial landmarks, action units, and physiological data, on which we achieved 94.95% and 94.91% of accuracy and F1-score, respectively.

Index Terms— Multimodal Learning, Transformer, Graph Convolutional Networks, Pain Detection.

1. INTRODUCTION

Pain is a complex and subjective human experience that profoundly affects our well-being. The ability to accurately detect and quantify pain has immense implications for various fields in healthcare, such as medical diagnosis, remote monitoring, sport medicine and rehabilitation. For instance, in sports medicine and rehabilitation settings, pain detection can assist in monitoring athletes' pain levels during training, competition, or recovery from injuries.

To design a robust pain detection model, it is crucial to consider multiple modalities, such as facial expressions, physiological indicators, and behavioral cues. As each modality is characterized by different statistical properties, it can be beneficial to explore the inner relationship between them. Multimodal machine learning (MML) is a hot multidisciplinary research field that involves the development of models that can extract and join information from multiple modalities.

The multimodal interactions among the intermediate representations of deep neural networks have led to very successful applications. Recent studies have shown that the intermediate representations of deep models could be as good or even better than using solely the last representation [8, 10, 14]. One core challenge in MML is the fusion of data from different modalities. Multimodal fusion strategy can be classified into three main categories: early, intermediate, and late fusion. In early fusion, the input modalities can be fused through concatenation, and the resulting feature vector is then treated like unimodal input. For intermediate fusion, higher-level representations from each input modality are learned through a stack of layers to discover within-modality correlations first and then are fused to discover cross-modality correlations. Regarding the late fusion strategy, different classifiers are trained on each modality and then an aggregation function is used to make the final decision.

In the last few years, the Transformer [12] model has been the de facto choice for dealing with natural language processing tasks [3]. In addition to language-related tasks, the Transformer has reached impressive results in many other areas, such as computer vision [4] and multimodal learning [1]. Recently, Graph Convolutional Networks have been widely used in the context of modeling relational data. Similar to Transformers, GCNs have shown successful results in multimodal applications [6].

Motivated by the vast success of the Transformers and GCN models for multimodal applications, we decided to investigate their relevance for the task of pain detection in multimodal settings. We propose a new fusion framework to explore the multimodal relations between the different levels of modality representations using a GCN. Our MMGT is built upon the intermediate Transformer layers representations of each modality. A graph is then constructed from these representations, where each node is connected to all other nodes within a modality and to nodes across other modalities corresponding to the same level of representation. To verify the effectiveness of our proposed approach, we have conducted ex-

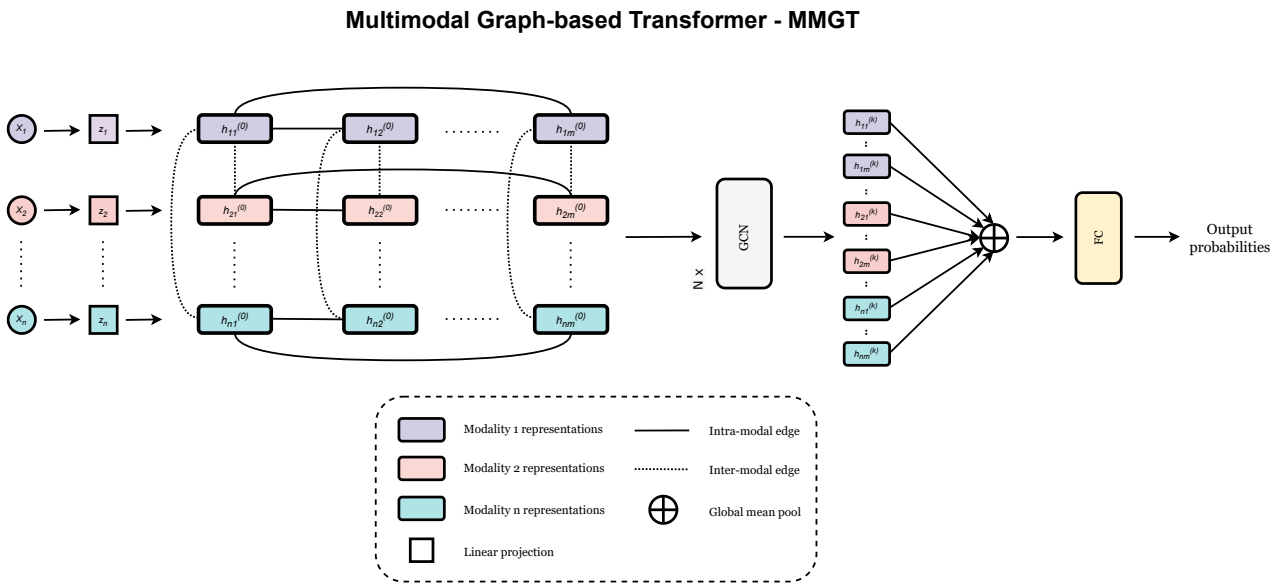

Fig. 1. Illustration of our MMGT framework, which is composed of two main building blocks: Unimodal Transformer encoder, and Multimodal Graph Convolutional Networks.

tensive experiments on the BP4D+ dataset [15]. Our MMGT model outperforms all existing approaches for the pain detection task.

The contributions of this work are threefold and can be summarized as follows: (1) the proposition of a new multimodal fusion framework that learns to combine the extracted representations from different modalities using a GCN; (2) to validate the proposed method and to further demonstrate the complementarity between the modalities, we provided a benchmark using single modalities and different combinations of two and three input modalities; (3) to the best of our knowledge, our MMGT is the first multimodal model trained on facial landmarks, action units, and physiological data.

2. PROPOSED APPROACH

This section introduces our MMGT model. Figure 1 shows an overview of the proposed framework, which is composed of two key components: (1) for each modality, a unimodal Transformer encoder extracts m intermediate representations; and (2) a GCN that learns to fuse the extracted representations.

2.1. Unimodal Transformer Encoder

The first stage of our framework consists of linearly projecting the data associated to each input modality $x_1, .., x_n$ to embedding vectors, $z_1, .., z_n$, each of dimension d_m. These projections are performed by the following learnable weight matrices $W_{x_1} \in \mathbb{R}^{d_{x_1} \times d_m}$, .., $W_{x_n} \in \mathbb{R}^{d_{x_n} \times d_m}$, where $d_{x_1}, .., d_{x_n}$ represent the dimension of each input modality. Then, we add positional encodings to each embedding vector to keep track of the relative order in the sequences. As the embeddings vectors and positional encodings have the same dimension, we can sum them up. Next, we feed the newly updated embedding vectors to unimodal Transformer encoders that will extract for each input modality a set of m intermediate representations. For a given input modality x_1, we obtain the following representations: $h_{11}, .., h_{1m}$.

2.2. Multimodal Fusion GCN

In order to capture the relationships between the extracted intermediate representations from the input modalities, we propose to use a spectral domain GCN.

2.2.1. Graph Construction

Our proposed multimodal framework is based on the construction of a graph $\mathcal{G} = (V, E)$ where V denotes the set of nodes initialized by the previously extracted representations, and E is the set of edges characterizing their relationships. We construct the graph as follows:
Nodes: every modality i is represented by a set of m nodes, initialized using their previously extracted representations $h_{i1}, .., h_{im}$. For n input modalities, we have $n \times m$ nodes.
Edges: every two nodes within a modality are connected together, and each node is also connected to nodes of other modalities corresponding to the same level of representation. We define the set of edges $E = E_{intra} \cup E_{inter}$ as the union of the sets of intra-modality edges, denoted by E_{intra}, and inter-modality edge, denoted by E_{inter}, defined as follows:

$$E_{intra} = \bigcup_{i=1}^{n} \bigcup_{j=1}^{m} \bigcup_{k=1}^{m} (h_{ij}, h_{ik}) \quad (1)$$

$$E_{inter} = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} \bigcup_{k=1}^{n} (h_{ji}, h_{ki}) \quad (2)$$

2.2.2. Graph Learning

We trained a spectral deep GCN based on the previously constructed graph $\mathcal{G}$. We define the graph convolution operator as in [9]:

$$\widetilde{H}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

where $\widetilde{A} = A + I_n$ denotes the adjacency matrix of the undirected graph $\mathcal{G}$ with inserted self-connections, I_n represents the identity matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ is the diagonal degree matrix, $W^{(l)}$ is a learnable weight matrix, and $\sigma(.)$ an activation function. H^l represents the matrix of activations in the l^{th} layer; $H^0 = X$, where X is the matrix of input node feature.

2.2.3. Pain Classifier

We initialize our graph nodes with the previously extracted representations from each modality. Let $h^0_{11}, .., h^0_{nm}$ be the set of initialized node representations of $\mathcal{G}$. We obtain $h^k_{11}, .., h^k_{nm}$ as the features encoded by the GCN after the k-th forward step.

$$h^k = [h^k_{11}, .., h^k_{1m}, .., h^k_{n1}, .., h^k_{nm}] \quad (3)$$

Then, we employed a global average pooling on the vector h^k, followed by a fully connected neural networks to predict the class label.

3. EXPERIMENTAL RESULTS

3.1. Datasets

We performed all our experiments on the BP4D+ [15] dataset, which includes 140 subjects performing a set of 10 tasks in order to elicit 10 authentic emotions. This dataset includes 3D face meshes, 2D RGB videos, thermal videos, facial landmarks (2D/3D/thermal), and eight physiological signals. Facial action units (AUs) were annotated for both the occurrence and intensity by FACS experts for four emotion categories (happiness, embarrassment, fear, and pain), on which only the most facially-expressive segments were encoded. In this study, we are only interested in those facially-expressive frames associated with the four previous emotions, as done in [5, 13].

Table 1. Unimodal pain detection: comparison with a state-of-the-art on the BP4D+ dataset.

	2D		3D		Thermal		AUs		Physio	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wu et al. [13]	91.59	89.46	91.27	89.30	83.53	83.37	-	-	**83.24**	**82.42**
Transformer	**92.99**	**92.93**	**92.45**	**92.15**	**86.81**	**85.41**	**92.11**	**92.15**	81.81	80.17

3.2. Data Processing

We first calculate the Euclidean distance between all peers of landmarks for each video frame using the provided facial landmarks (2D/3D/thermal), as it gives better results than using the raw landmarks. Next, physiological data were downsampled to the same frequency as the one associated with the video recordings (25 fps). Afterward, for each modality, we generated non-overlapping sliding windows of dimension 350. If the length of the data sequence is less than 350, we pad the sequence with the data associated with the last frame.

3.3. Results

We carried out unimodal and multimodal pain detection experiments on the BP4D+ dataset. In order to perform pain detection as a binary classification among all four emotion retained, we treat the pain sequences as positive classes, and the remaining three as negative classes. Following prior works [13], we employed a subject-independent 10-fold cross-validation strategy. For evaluation, we used the accuracy and weighted average F1-score.

3.3.1. Unimodal

In Table 1, we show the performances of the Transformer model and those from a state-of-the-art method [13] for the task of pain detection using the following modalities: physiological data, 2D, 3D, and thermal facial landmarks. The Transformer model outperformed the method proposed by [13] in terms of accuracy and F1-score when using 2D, 3D, and thermal facial landmarks. Specifically, the Transformer model achieved a 1.40% and 3.47% improvement in terms of accuracy and F1-score, respectively, over the previous method for 2D landmarks, a 1.18% and 2.05% improvement for 3D landmarks, and a 3.28% and 2.04% improvement for thermal landmarks.

To further demonstrate the relevance of Transformers for multimodal applications, we evaluate our MMGT framework which, as explained earlier, leverages multimodal interactions between intermediate modality representations.

3.3.2. Multimodal and Ablation Study

Our experiments with multimodal data are summarized in Tables 2 and 3, on which we evaluate our proposed MMGT

Table 2. Multimodal pain detection: comparison with a state-of-the-art on the BP4D+ dataset, and ablation study on input modalities, fusion techniques and model settings.

	Two modalities								Three modalities					
	2D + Physio		3D + Physio		Thermal + Physio		AUs + Physio		2D + AUs + Physio		3D + AUs + Physio		Thermal + AUs + Physio	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wu et al. [13]	93.45	91.37	92.66	90.47	89.07	88.96	-	-	-	-	-	-	-	-
MMT-early	92.65	92.63	90.99	90.74	85.67	84.49	90.79	90.43	93.31	93.33	92.03	92.01	92.42	92.23
MMT-inter	90.82	90.77	88.51	88.27	79.14	80.15	**94.06**	94.01	93.17	93.31	93.62	93.59	93.66	93.67
MMT-late	91.02	91.04	87.39	87.79	77.69	78.92	93.89	93.97	93.64	93.79	92.76	92.92	93.62	93.74
MMT-all	93.53	93.38	91.55	91.19	86.35	85.78	93.70	93.74	94.09	94.00	93.66	93.53	93.39	93.43
MMGT (ours)	**93.90**	**93.82**	**93.72**	**93.59**	**89.45**	**89.14**	**94.07**	**94.10**	**94.95**	**94.91**	**94.41**	**94.31**	**93.87**	**93.93**

Table 3. Comparison of our pain detection method with state-of-the-art results.

Method	Acc	F1 (pain class)
AUs + Physio		
Hinduja et al. [5]	89.20	75.00
MMGT (Ours)	**94.07**	**88.33**
Method	**Acc**	**F1**
2D + Physio		
Szczapa et al. (late fusion) [11]	82.77	76.32
Szczapa et al. (early fusion) [11]	84.32	78.83
Huang et al. (early fusion) [7]	87.94	87.16
Huang et al. (late fusion) [7]	89.36	89.13
Choo et al. (late fusion) [2]	89.08	88.68
Choo et al. (early fusion) [2]	89.80	89.46
Wu et al. [13]	93.45	91.37
MMGT (Ours)	**93.90**	**93.82**

model against machine learning [5], deep learning [2, 7], and geometric-based models [11, 13]. The following pain recognition approaches [2, 7, 11] have been reimplemented by [13] and the scores reported in Table 3 are from [13]. We trained our models on different combinations of the modalities given in Table 1. We first combined physiological data with all other four modalities, as shown in Table 2. Then, for the combinations of three modalities, we combined AUs and physiological data with all types of facial landmarks.

Using two modalities, MMGT achieve state-of-the-art results on all tested combinations. Compared to Wu et al. [13], the largest improvements are observed for the combination of 3D landmarks and physiological data, on which our MMGT model outperformed the later by 3.12% in terms of F1-score. Among all tested combinations of two modalities, our best results were obtained using AUs and physiological data, on which our model achieved 94.07% and 94.10% of accuracy and F1-score, respectively. Compared to Hinduja et al. [5], the only state-of-the-art method which also combines AUs and physiological data, we observe an improvement of 4.87% and 13.33% in terms of accuracy and F1-scores, respectively.

It is worth mentioning that, for a fair comparison with [5], we reported the F1-score only for the pain class in Table 3. Whereas, to provide a fair comparison with the other models, the remaining F1-scores reported in Tables 2 and 3 are the weighted average F1-score for the pain and non-pain classes.

Given that our best combination of two modalities were achieved with AUs and physiological data, we tested if we could further improve the MMGT results by considering a third modality, leading to the combination of AUs, physiological data and each one of the landmark data (2D, 3D and thermal). As shown in Table 3, our hypothesis was validated in the cases where 2D and 3D landmarks were considered as a third modality, but not when thermal landmarks were included. Moreover, the best results among all tested methods were obtained with MMGT trained on 2D landmarks, AUs and physiological data. The latter achieved respectively 94.95% and 94.91% of accuracy and F1-score, establishing a new state-of-the-art for the pain detection task on BP4D+.

As part of our ablation study, we observed that fusing modalities representations with GCNs yields the highest evaluation scores when compared to traditional fusion methods, such as early, intermediate and late fusion using Transformers. In Table 2, we present our evaluation scores obtained with the aforementioned fusion techniques for different combinations of two and three modalities. More precisely, MMT-early concatenates at the input level the different modalities, MMT-inter concatenates the final representation layer of each Transformer, and MMT-late aggregates the final decision of each Transformer. As we can see in Table 2, MMGT (the only graph based model) outperformed all fusion techniques on the task of pain detection for all combination of two and three modalities. We believe that these positive results are mainly attributed to two components: (1) the use of intermediate Transformer representations, and (2) the use of a graph to facilitate their fusion. Table 2 shows that, in general, traditional fusion techniques result in inferior performance compared to the best-performing modality used individually, which does not hold true for MMGT. For example, fusing physiological data with 2D landmarks using traditional fusion techniques leads to a drop of at least 0.30% in terms of F1-score com-

pared to 2D landmarks used alone (see Table 1). On the other hand, MMGT improves upon all individual modalities.

Finally, to conduct an ablation study on the use of GCNs to efficiently combine the intermediate representations of multiple modalities, we compare MMGT with a multimodal Transformer (MMT-all) which does not make use of graphs. More exactly, MMT-all directly concatenates the intermediate representations $h^0_{11}, .., h^0_{nm}$, which are then fed into a fully-connected layer for final decision. As shown in Table 2, MMGT outperforms MMT-all for all combinations of two and three modalities. For instance, when using 3D landmarks and physiological data, we remarked a difference of 2.17% and 2.40% in accuracy and F1-score, respectively.

Those results bring to light several aspects of MMGT multimodal learning. First, the complementarity between different modalities (*e.g.* adding physiological data with visual features consistently enhances pain detection performance). Second, the successive addition of modalities nearly always lead to better classification results. Third, the way of leveraging the multimodal interactions between different learned representations has a high impact on the final task at hand.

4. CONCLUSION

In this study, we tackled the problem of multimodal learning for pain detection. We propose a new multimodal fusion framework called Multimodal Graph-based Transformer (MMGT), which employs a graph to capture the multimodal relations between the intermediate representations extracted from unimodal Transformers. We conducted extensive experiments on the BP4D+ dataset, on which we established a new state-of-the-art for pain detection. Our best results were obtained using 2D facial landmarks, AUs, and physiological data, on which we achieved 94.95% and 94.91% of accuracy and F1-score, respectively. In a future work, we plan to investigate our proposed framework on other multimodal datasets.

5. REFERENCES

[1] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 34:24206–24221, 2021.

[2] K. W. Choo and T. Du. Pain detection from facial landmarks using spatial-temporal deep neural network. In *ICDIP 2021*, volume 11878, pages 593–597. SPIE.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] S. Hinduja, S. Canavan, and G. Kaur. Multimodal fusion of physiological signals and facial action units for pain recognition. In *IEEE FG*, pages 577–581, 2020.

[6] J. Hu, Y. Liu, J. Zhao, and Q. Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*, 2021.

[7] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng. Hybnet: a hybrid network structure for pain intensity estimation. *The Visual Computer*, 38(3):871–882, 2022.

[8] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *IEEE CVPR*, pages 13289–13299, 2020.

[9] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[10] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *IEEE ICASSP*, pages 6419–6423, 2020.

[11] B. Szczapa, M. Daoudi, S. Berretti, P. Pala, A. Del Bimbo, and Z. Hammal. Automatic estimation of self-reported pain by interpretable representations of motion dynamics. In *IEEE ICPR*, pages 2544–2550, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[13] Y. Wu, M. Daoudi, A. Amad, L. Sparrow, and F. D'Hondt. Fusion of physiological and behavioural signals on spd manifolds with application to stress and pain detection. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2949–2955. IEEE, 2022.

[14] Y. Wu, Z. Zhang, P. Peng, Y. Zhao, and B. Qin. Leveraging multi-modal interactions among the intermediate representations of deep transformers for emotion recognition. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 101–109, 2022.

[15] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *IEEE CVPR*, 2016.