Evaluating the Effectiveness, Generalizability, and Explainability of Video Swin Transformers on Automated Pain Detection

Maximilian Rau[1], supervised by Itir Onal Ertugrul[1]

Utrecht University, Utrecht, The Netherlands

Abstract. Recent advancements in computer vision, particularly with transformer-based models, have shown promise for automated pain assessment using facial expressions. This study evaluates the Video Swin Transformer (VST), which leverages temporal dynamics for nuanced pain detection. We compare the VST's performance to other transformer models, like Swin Transformer and Vision Transformer (ViT). Moreover, we also demonstrate that higher temporal depth improves performance. While the use of Focal Loss to address class imbalance in the UNBC McMaster dataset was ineffective, our research also highlights the importance of diverse datasets for model generalizability. Attention map analysis confirms that the VST models focus on pain-related facial regions, enhancing their explainability. The VST-0 and VST-1-TD models achieved new state-of-the-art F1 scores of 0.56 ± 0.06 and 0.59 ± 0.04, respectively, and competitive AUC scores. This work emphasizes the VST architecture's potential in both automated pain assessment and broader facial expression analysis. Code is available at `https://github.com/MRausus/VST-APA`

Keywords: Automated Pain Detection · Video Swin Transformer · Generalizability · Explainability · Focal Loss.

1 Introduction

Pain assessment is critical in clinical diagnostics and treatment. Automated pain detection using facial expressions has gained traction due to its non-invasive nature and reliability as a pain indicator [10][12][9]. Recent advancements in computer vision have seen a shift from Convolutional Neural Networks (CNNs) to transformer-based models like ViTs [2], which set new benchmarks in various vision tasks. However, ViTs have limitations which have been mitigated using Swin Transformers. This research introduces the use of VST [5] to model spatiotemporal dynamics of facial expressions, offering an advanced approach for automated pain detection.

2 Methodology

The proposed methodology involves using the VST [2] to detect pain through facial expressions by using spatiotemporal dynamics. The models were trained

and evaluated using two datasets: the UNBC McMaster Shoulder Pain dataset [7] and the BioVid Heat Pain Database [13]. However, the BioVid dataset was used exclusively for cross-dataset validation as a test set. Preprocessing steps included 2D alignment of faces with RetinaFace [1], normalization, and subsequence generation. We employ a five-fold cross-validation strategy, dividing the UNBC database into five subsets on patient level with similar class distributions and training models on different combinations of these subsets. Models were trained using a combination of cross-entropy loss with oversampling for handling class imbalance. Evaluation employed both quantitative (F1 score, AUC) and qualitative (attention visualization maps [8]) methods to assess model effectiveness, generalizability, and explainability.

3 Experiments and Results

Several experiments were conducted using the UNBC McMaster to evaluate the VST's performance compared to the Swin Transformer (ST-0) and Vision Transformer (ViT-0). A performance comparison of our models and previous work is presented in Table 1. The VST-0 model with four frames achieved an F1 score of 0.56, while the VST-1-TD model with eight frames achieved a higher F1 score of 0.59, highlighting the benefits of incorporating more temporal information. Another ablation study including the VST-2-FL model showed that Focal Loss [4] as an alternative to handle class imbalance is not sufficient in our scenario. In cross-dataset validation with the BioVid dataset, the VST-0 model demonstrated the best generalizability regarding F1, although all models struggled to maintain high accuracy. Attention heatmaps (see Figure 1) confirmed that VST models focus on relevant facial regions associated with pain, such as the eyes and nose, validating their effectiveness and explainability.

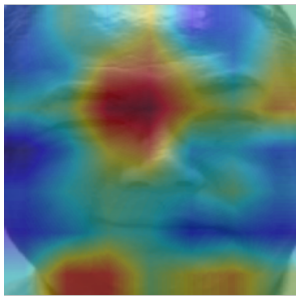

Fig. 1: Attention heatmap obtained by averaging the heatmaps of 100 true positive images using VST-0. Red areas indicate high attention, while blue areas indicate low attention.

Table 1: Performance comparison of our models with previous work

Model	**F1 score**	**AUC**
CDL [11]	0.46 ± 0.18	–
PFDL [11]	0.47 ± 0.20	–
SPTS + C [7]	–	0.84
SPTS + S + C [6]	–	0.85
ViT-1-D [3]	0.55 ± 0.15	**0.88**
ViViT-1-D [3]	0.55 ± 0.13	0.86
VST-0	0.56 ± 0.06	0.85
ST-0	0.53 ± **0.04**	0.80
ViT-0	0.55 ± 0.09	0.87
VST-1-TD	**0.59** ± **0.04**	0.87
VST-2-FL	0.44 ± 0.11	0.77

References

1. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild (2019)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
3. Fiorentini, G., Ertugrul, I.O., Salah, A.A.: Fully-attentive and interpretable: vision and video vision transformers for pain detection (2022)
4. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
5. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer (2021)
6. Lucey, P., Cohn, J., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., Prkachin, K.: Automatically detecting pain in video through facial action units. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society **41**, 664–74 (11 2010). https://doi.org/10.1109/TSMCB.2010.2082525
7. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: The unbc-mcmaster shoulder pain expression archive database. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG). pp. 57–64 (2011). https://doi.org/10.1109/FG.2011.5771462
8. Nguyen, H.C., Lee, H., Kim, J.: Inspecting explainability of transformer models with additional statistical information (2023)
9. Prkachin, K.M.: The consistency of facial expressions of pain: A comparison across modalities. Pain **51**(3), 297–306 (1992). https://doi.org/10.1016/0304-3959(92)90213-u
10. 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). https://doi.org/10.1016/j.pain.2008.04.010
11. Rudovic, O., Tobis, N., Kaltwang, S., Schuller, B., Rueckert, D., Cohn, J.F., Picard, R.W.: Personalized federated deep learning for pain estimation from face images. arXiv preprint arXiv:2101.04800 (2021), https://arxiv.org/abs/2101.04800
12. Sheu, E., Versloot, J., Nader, R., Kerr, D., Craig, K.D.: Pain in the elderly. The Clinical Journal of Pain **27**(7), 593–601 (2011). https://doi.org/10.1097/ajp.0b013e31820f52e1
13. Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H.C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A.O., Moreira da Silva, G.: 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). https://doi.org/10.1109/CYBConf.2013.6617456