




Article

Two-Stream Attention Network for Pain Recognition from Video Sequences

Patrick Thiam ^{1,2} , Hans A. Kestler ^{1,†}  and Friedhelm Schwenker ^{2,*,†} 

¹ Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany; patrick.thiam@uni-ulm.de (P.T.); hans.kestler@uni-ulm.de (H.A.K.)

² Institute of Neural Information Processing, Ulm University, James-Frank-Ring, 89081 Ulm, Germany

* Correspondence: friedhelm.schwenker@uni-ulm.de

† Equally contributing senior authors.

Received: 23 January 2020; Accepted: 2 February 2020; Published: 4 February 2020



Abstract: Several approaches have been proposed for the analysis of pain-related facial expressions. These approaches range from common classification architectures based on a set of carefully designed handcrafted features, to deep neural networks characterised by an autonomous extraction of relevant facial descriptors and simultaneous optimisation of a classification architecture. In the current work, an end-to-end approach based on attention networks for the analysis and recognition of pain-related facial expressions is proposed. The method combines both spatial and temporal aspects of facial expressions through a weighted aggregation of attention-based neural networks' outputs, based on sequences of Motion History Images (MHIs) and Optical Flow Images (OFIs). Each input stream is fed into a specific attention network consisting of a Convolutional Neural Network (CNN) coupled to a Bidirectional Long Short-Term Memory (BiLSTM) Recurrent Neural Network (RNN). An attention mechanism generates a single weighted representation of each input stream (MHI sequence and OFI sequence), which is subsequently used to perform specific classification tasks. Simultaneously, a weighted aggregation of the classification scores specific to each input stream is performed to generate a final classification output. The assessment conducted on both the *BioVid Heat Pain Database (Part A)* and *SenseEmotion Database* points at the relevance of the proposed approach, as its classification performance is on par with state-of-the-art classification approaches proposed in the literature.

Keywords: convolutional neural networks; long short-term memory recurrent neural networks; information fusion; pain recognition

1. Introduction

An individual's affective disposition is often expressed throughout facial expressions. Human beings are therefore able to assess someone's current mood or state of mind by observing his or her facial demeanour. Therefore, an analysis of facial expressions can provide some valuable insight about one's emotional and psychological state. Thus, facial expression recognition (FER) has been attracting a lot of interest from the research community in the recent decades and constitutes a steadily growing area of research, particularly in the domains of machine learning and computer vision. The current work focuses on the analysis of facial expressions for the assessment and recognition of pain in video sequences. More specifically, a two-stream attention network is designed, with the objective of combining both temporal and spatial aspects of facial expressions, based on sequences of motion history images [1] and optical flow images [2], to accurately discriminate between neutral, low, and high levels of nociceptive pain. The current work is organised as follows. An overview of pain recognition approaches based on facial expressions is provided in Section 2, followed by a thorough

description of the proposed approach in Section 3. In Section 4, a description of the datasets used for the assessment of the proposed approach as well as the performed experiments is provided, followed by a description of the corresponding results. The current work is subsequently concluded in Section 5 with a short discussion and description of potential future works.

2. Related Work

Recent advances in both domains of computer vision and machine learning, combined with the release of several datasets designed specifically for pain-related research (e.g., *UNBC-McMaster Shoulder Pain Expression Archive Database* [3], *BioVid Heat Pain Database* [4], *Multimodal EmoPain Database* [5] and *SenseEmotion Database* [6]), have fostered the development of a multitude of automatic pain assessment and classification approaches. These methods range from unimodal approaches, characterised by the optimisation of an inference model based on one unique and specific input signal (e.g., video sequences [7,8], audio signals [9,10] and bio-physiological signals [11–13]), to multimodal approaches that are characterised by the optimisation of an information fusion architecture based on parameters stemming from a set of distinctive input signals [14–16].

Regarding pain assessment based on facial expressions, several approaches have been proposed, ranging from conventional supervised learning techniques based on specific sets of handcrafted features, to deep learning techniques. These approaches rely on an effective preprocessing of the input signal (which in this case consists of a set of images or video sequences) and involves the localisation, alignment and normalisation of the facial area in each input frame. Moreover, further preprocessing techniques include the localisation and extraction of several fiducial points characterising specific facial landmarks, and in some cases, the continuous extraction of facial Action Units (AUs) [17,18]. The preprocessed input signal, as well as the extracted parameters, are subsequently used to optimise a specific inference model based on different methods. In [19], the authors use an ensemble of linear Support Vector Machines (SVMs) [20] (each trained on a specific AU), in which inference scores are subsequently combined using Logistical Linear Regression (LLR) [21] for the detection of pain at a frame-by-frame level. The authors in [22] apply a *k*-Nearest Neighbours (*k*-NN) [23] model on geometric features extracted from a specific set of facial landmarks for the recognition of AUs. Subsequently, the pain intensity in a particular frame is evaluated based on the detected AUs by using a pain evaluation scale provided by Prkachin and Solomon [24]. Most recently, the authors in [25] improve the performance of a pain detection system based on automatically detected AUs by applying a transfer learning approach based on neural networks to map automated AU codings to a subspace of manual AU codings. The encoded AUs are subsequently used to perform pain classification, using an Artificial Neural Network (ANN) [26].

In addition to AU-based pain assessment approaches, several techniques based on either facial texture, shape, appearance and geometry or on a combination of several of such facial descriptors have been proposed. Yang et al. [27] assess several appearance-based facial descriptors by comparing the pain classification performance of each feature with its spatio-temporal counterpart using SVMs. The assessed spatial descriptors consist of Local Binary Patterns (LBP) [28], Local Phase Quantization (LPQ) [29], Binarized Statistical Image Features (BSIF) [30] as well as each descriptor's spatio-temporal counterpart extracted from video sequences on three orthogonal planes (LBP-TOP, LPQ-TOP and BSIF-TOP). In [8,31], the authors propose several sets of spatio-temporal facial action descriptors based on both appearance- and geometry-based features extracted from both the facial area, as well as the head pose. Those descriptors are further used to perform the classification of several levels of pain intensities using a Random Forest (RF) [32] model. Similarly, the authors in [7,14,15,33], propose several spatio-temporal descriptors extracted either from the localised facial area or from the estimated head pose, including, among others, Pyramid Histograms of Oriented Gradients (PHOG) [34] and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [35], to perform the classification of several levels of nociceptive pain. The classification experiments are also performed with RF models and ANNs.

Alongside handcrafted feature-based approaches, several techniques based on deep neural networks have also been proposed for the assessment of pain induced facial expressions. Such approaches are characterised by the joint extraction of relevant descriptors (from the preprocessed raw input data) and optimisation of an inference model, based on neural networks in an end-to-end manner. In [36–38], the authors propose a hybrid deep neural network pain detection architecture characterised by the combination of a feature embedding network consisting of a Convolutional Neural Network (CNN) [39] with a Long Short-Term Memory (LSTM) [40] Recurrent Neural Network (RNN), to take advantage of both spatial and temporal aspects of facial pain expressions in video sequences. Soar et al. [41] propose a similar approach by combining a CNN with a Bidirectional LSTM (BiLSTM), and using a Variable-State Latent Conditional Random Field (VRS-CRF) [42] instead of a conventional Multi-Layer Perceptron (MLP) to perform the classification. In [43], the authors also use a similar hybrid approach as in [36,37]; however, in this case, the feature embedding CNN is coupled to two distinct LSTM networks. The outputs of the LSTM networks are further concatenated and a MLP is used to perform the classification of the pain intensities in video sequences. Furthermore, Zhou et al. [44] propose a Recurrent Convolutional Neural Network (RCNN) [45] architecture for the continuous estimation of pain intensity in video sequences at the frame-level, whereas Wang et al. [46] propose a transfer learning approach, consisting of the regularisation of a face verification network, which is subsequently applied to a pain intensity regression task.

The current work focuses on the analysis of facial expressions for the discrimination of neutral, low and high levels of nociceptive pain in video sequences. Thereby, an end-to-end hybrid neural network characterised by the integration of spatial and temporal information at both the representational level of the input data (OFI and MHI) and the structural level of the proposed architecture (hybrid CNN-BiLSTM) is proposed. Furthermore, frame attention parameters [47] are integrated into the proposed architecture to generate an aggregated representation of the input data based on an estimation of the representativeness of each single input frame, in relation with the corresponding level of nociceptive pain. An extensive assessment of the proposed architecture is performed on both *BioVid Heat Pain Database (Part A)* [4] and *SenseEmotion Database* [6].

3. Proposed Approach

A video sequence can be characterised by both its spatial and temporal components. The spatial component represents the appearance (i.e., texture, shape and form) of each frame's content, whereas the temporal component represents the perceived motion across consecutive frames due to dynamic changes of the content's appearance through time. Most of the deep neural network approaches designed for the assessment of pain-related facial expressions generate spatio-temporal descriptors of the input data in two distinct and conjoint stages: a specific feature embedding neural network (which is commonly a pre-trained CNN) first extracts appearance based descriptors from the individual input frames (which are greyscale or colour images), and a recurrent neural network is subsequently used for the integration of the input's temporal aspect based on sequences of previously extracted appearance features, thus generating spatio-temporal representations of video sequences that are used for classification or regression tasks. Therefore, both the temporal and spatial aspects of video sequences are integrated uniquely at the structural level (e.g., the architecture of the neural network) of such approaches. The current approach extends this specific technique by additionally integrating motion information at the representational level (e.g., input data) of the architecture throughout sequences of motion history images [1] and optical flow images [2].

3.1. Motion History Image (MHI)

Introduced by Bobick and Davis [48], a MHI consists of a scalar-valued image depicting both the location and direction of motion in a sequence of consecutive images, based on the changes of pixel intensities of each image through time. The intensity of a pixel in a MHI is a function of the temporal

motion history at that specific point. A MHI H_τ is computed using an update function $\Psi(x, y, t)$, and is defined as follows,

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } \Psi(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - \delta) & \text{otherwise} \end{cases} \quad (1)$$

where (x, y) represents the pixel's location, t the time and τ the temporal extent of the observed motion (e.g., the length of a sequence of images); $\Psi(x, y, t) = 1$ is synonym of motion at the location (x, y) and at the time t ; and δ represents a decay parameter. The update function $\Psi(x, y, t)$ is defined as follows,

$$\Psi(x, y, t) = \begin{cases} 1 & \text{if } D(x, y, t) \geq \zeta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where ζ is a threshold; $D(x, y, t)$ represents the absolute value of the difference of pixel intensity values of consecutive frames and is defined as follows,

$$D(x, y, t) = |I(x, y, t) - I(x, y, t \pm \Delta t)| \quad (3)$$

where $I(x, y, t)$ represents the pixel intensity at the location (x, y) and at the time t ; Δt represents the temporal distance between the frames.

Therefore, the computation of a MHI consists in first performing image differencing [49] between a specific, preceding frame and the current t th frame, and detecting the pixel locations where a substantial amount of movement has occurred (depending on the value assigned to the threshold ζ) based on Equation (2). Subsequently, Equation (1) is used to assign pixel values to the MHI. If a motion has been detected at the location (x, y) of the t th frame, a pixel value of τ is assigned at that location. Otherwise, the previous pixel value of that location is reduced by δ , thereby indicating the temporal occurrence of the motion at that specific location, according to the actual time t . This whole process is conducted iteratively until the entire sequence of images has been processed. The temporal history of motion is thereby encoded into the resulting MHI. Therefore, a whole sequence of images can be encoded into a single MHI. However, in the current work, a sequence of MHIs is generated from each single sequence of images by saving each single MHI generated at each single step of the iterative process described earlier. The resulting sequence of images is used as input for the designed deep neural networks. A depiction of such a sequence of MHIs can be seen in Figure 1b, with the corresponding sequence of greyscale images depicted in Figure 1a.

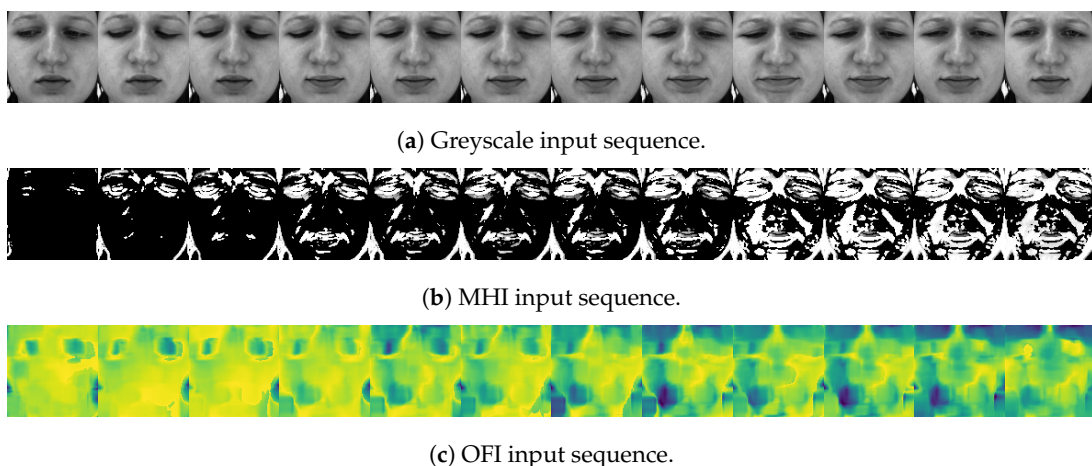


Figure 1. Data preprocessing. Following the detection, alignment, normalisation and extraction of the facial area in each frame of a video sequence, the images are converted into greyscale. MHI and OFI sequences are subsequently generated.

3.2. Optical Flow Image (OFI)

Optical flow refers to the apparent motion of visual features (e.g., corners, edges, textures and pixels) in a sequence of consecutive images. It is characterised by a set of vectors (optical flow vectors) defined either at each location (x, y) of an entire image (dense optical flow [50,51]), or at specific locations of a predefined set of visual features (sparse optical flow [2,52]). The orientation of an optical flow vector depicts the direction of the apparent motion, whereas the magnitude of an optical flow vector depicts the velocity of the apparent motion of the corresponding visual feature in consecutive frames. Thus, an OFI provides a compact description of the location, direction and velocity of a specific motion occurring in consecutive frames. The estimation of the optical flow is based on the brightness constancy assumption, which stipulates that pixel intensities are constant between consecutive frames. If $I(x, y, t)$ is the pixel intensity at the location (x, y) and at the time t , the brightness constancy assumption can be formulated as follows,

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \quad (4)$$

where $(\delta x, \delta y, \delta t)$ represents a small motion. By applying a first-order Taylor expansion, $I(x + \delta x, y + \delta y, t + \delta t)$ can be written as follows,

$$I(x + \delta x, y + \delta y, t + \delta t) \approx I(x, y, t) + \frac{\partial I}{\partial x} \delta x + \frac{\partial I}{\partial y} \delta y + \frac{\partial I}{\partial t} \delta t. \quad (5)$$

Thus,

$$\frac{\partial I}{\partial x} \delta x + \frac{\partial I}{\partial y} \delta y + \frac{\partial I}{\partial t} \delta t \approx 0 \quad (6)$$

and by dividing each term by δt , the optical flow constraint equation can be written as follows,

$$\frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{dI}{dt} \approx 0. \quad (7)$$

Resolving the optical flow constraint equation (Equation (7)) consists of the estimation of both parameters $u = \frac{dx}{dt}$ and $v = \frac{dy}{dt}$. Several methods have been proposed to perform this specific task. The authors in [53,54] propose an overview of such approaches. In the current work, dense optical flow is applied, using the method of Farnebäck [50], to compute OFIs from consecutive greyscale images. A depiction of such a sequence of images can be seen in Figure 1c (both motion direction and motion velocity are color encoded).

3.3. Network Architecture

As opposed to still images, the motion component of a video sequence is integrated into both MHIs and OFIs, therefore providing more valuable information for facial expressions analysis. Therefore, the proposed architecture consists of a multi-view learning [55] neural network with both OFIs and MHIs as input channels. An overall illustration of the proposed two-stream neural network can be seen in Figure 2. In a nutshell, an attention network specific to each input channel (OFIs and MHIs) first generates a weighted representation from the j th input sequence (h_j^{ofi} and h_j^{mhi}). The generated representation is subsequently fed into a channel specific classification model (which in this case is a MLP). The resulting class probabilities of each channel ($score_j^{ofi}$ and $score_j^{mhi}$) are further fed into an aggregation layer with a linear output function, where a weighted aggregation of the provided scores is performed as follows,

$$score_j = \alpha_{ofi} \cdot score_j^{ofi} + \alpha_{mhi} \cdot score_j^{mhi} \quad (8)$$

with the constraint

$$\alpha_{ofi} + \alpha_{mhi} = 1. \quad (9)$$

The entire architecture is trained in an end-to-end manner by using the following loss function,

$$\mathcal{L} = \lambda_{ofi} \cdot \mathcal{L}_{ofi} + \lambda_{mhi} \cdot \mathcal{L}_{mhi} + \lambda_{agg} \cdot \mathcal{L}_{agg} \quad (10)$$

where the loss functions of each input channel and of the aggregation layer are respectively depicted with \mathcal{L}_{ofi} , \mathcal{L}_{mhi} and \mathcal{L}_{agg} . The parameters λ_{ofi} , λ_{mhi} and λ_{agg} correspond to the regularisation parameters of each respective loss function. Once the network has been trained, unseen samples are classified based on the output of the aggregation layer.

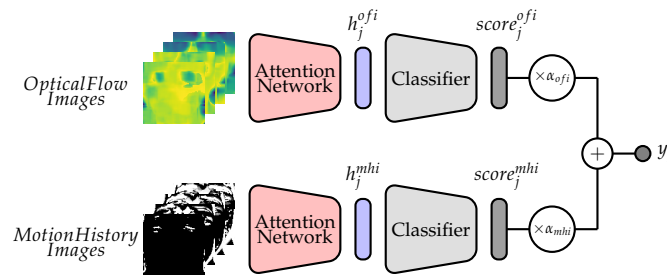


Figure 2. Two-Stream Attention Network with Weighted Score Aggregation.

The attention network (see Figure 3) consists of a CNN coupled to a BiLSTM with a frame attention module [47]. The CNN consists of a time distributed feature embedding network which takes a single facial image $im_{k,j}$ as input and generates a fixed-dimension feature representation $X_{k,j}$ specific to that image. Therefore, the output of the j th input sequence of facial images $\{im_{k,j}\}_{k=1}^l$ consists of a set of facial features $\{X_{k,j}\}_{k=1}^l$. The temporal component of the sequence of images is further integrated by using a BiLSTM layer. A BiLSTM [56] RNN is an extension of a regular LSTM [40] RNN, to enable the use of context representations in both forward and backward directions.

It consists of two LSTM layers, one processing the input sequence $\{X_{1,j}, X_{2,j}, \dots, X_{l,j}\}$ sequentially forward in time (from $X_{1,j}$ to $X_{l,j}$) and the second processing the input sequence sequentially backward in time (from $X_{l,j}$ to $X_{1,j}$). A LSTM RNN is capable of learning long-term dependencies in sequential data, while avoiding the vanishing gradient problem of standard RNNs [57]. This is achieved throughout the use of cell states (see Figure 4), which regulate the amount of information flowing through a LSTM network throughout the use of three principal gates: forget gate (f_t), input gate (i_t) and output gate (o_t). The cell's output h_t (at each time step t) is computed, given a specific input x_t , the previous hidden state h_{t-1} , and the previous cell state C_{t-1} , as follows,

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (11)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (12)$$

$$\tilde{C}_t = \tanh(W_c s_t + U_c h_{t-1} + b_c) \quad (13)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \quad (14)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (15)$$

$$h_t = o_t \otimes \tanh(C_t) \quad (16)$$

where σ represents the sigmoid activation function $\sigma(x) = (1 + \exp(-x))^{-1}$ and \tanh represents the hyperbolic tangent activation function. The element-wise multiplication operator is represented by the symbol \otimes . The weight matrices for the input x_t are represented by W_i , W_f , W_o and W_c for the input gate, forget gate, output gate and cell state, respectively. Analogously, the weight matrices for the previous

hidden state h_{t-1} for each gate are represented by U_i, U_f, U_o and U_c . The amount of information to be further propagated into the network is controlled by the forget gate (Equation (11)), the input gate (Equation (12)) and the computed cell state candidate \tilde{C}_t (Equation (13)). These parameters are subsequently used to update the cell state C_t based on the previous cell state C_{t-1} (Equation (14)). The output of the cell can subsequently be computed using both Equation (15) and Equation (16).

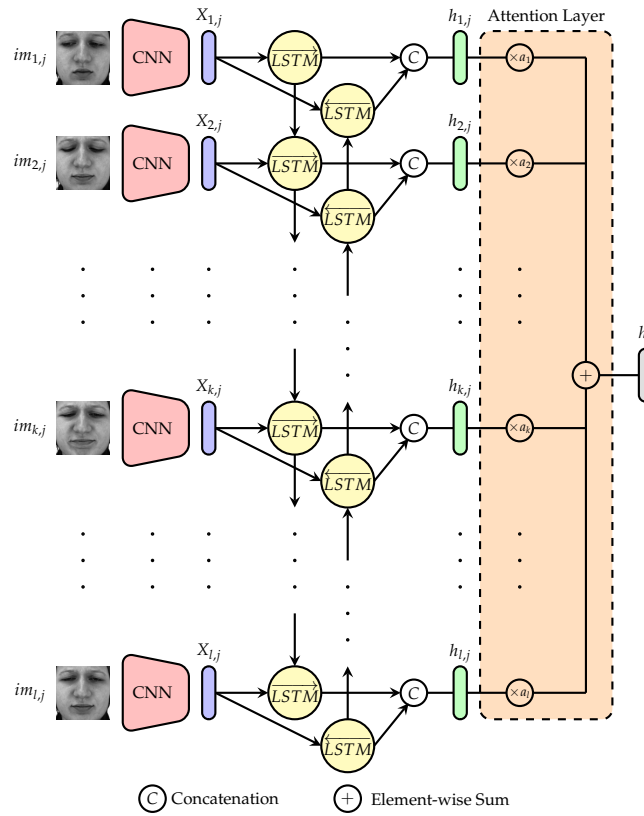


Figure 3. Attention Network.

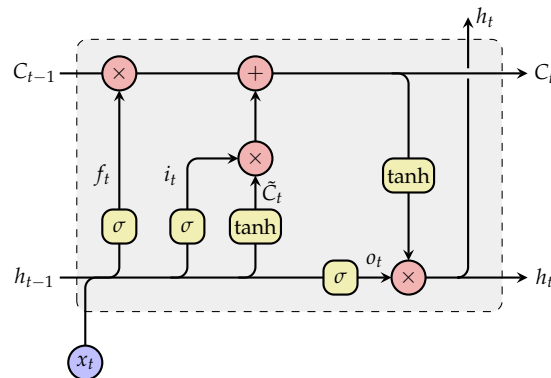


Figure 4. Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN).

In the current work, the hidden representation stemming from the forward pass $\{\vec{h}_{1,j}, \vec{h}_{2,j}, \dots, \vec{h}_{l,j}\}$ and the one stemming from the backward pass $\{\overleftarrow{h}_{1,j}, \overleftarrow{h}_{2,j}, \dots, \overleftarrow{h}_{l,j}\}$ are subsequently concatenated $\{[\vec{h}_{1,j}, \overleftarrow{h}_{1,j}], [\vec{h}_{2,j}, \overleftarrow{h}_{2,j}], \dots, [\vec{h}_{l,j}, \overleftarrow{h}_{l,j}]\}$ and fed into the next layer. For the sake of simplicity, the output of the BiLSTM layer will be depicted as follows, $\{h_{1,j}, h_{2,j}, \dots, h_{l,j}\}$ (with $h_{k,j} = [\vec{h}_{k,j}, \overleftarrow{h}_{k,j}]$). The next layer consists of an attention layer, where self-attention weights $\{a_k\}_{k=1}^l$ are

computed and subsequently used to generate a single weighted representation of the input sequence. The self-attention weights are computed as follows,

$$\alpha_k = \text{elu} \left(W_k h_{k,j} + b_k \right) \quad (17)$$

$$a_k = \frac{\exp(\alpha_k)}{\sum_{i=1}^l \exp(\alpha_i)} \quad (18)$$

where W_k are the weights specific to the input feature representation $h_{k,j} = \left[\overrightarrow{h_{k,j}}, \overleftarrow{h_{k,j}} \right]$ and elu represents the exponential linear unit activation function [58], which is defined as

$$\text{elu}_\alpha(x) = \begin{cases} \alpha \cdot (\exp(x) - 1) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (19)$$

with $\alpha = 1$. Each self-attention weight expresses the relevance of a specific image for the corresponding emotional state expressed within the video sequence. Thereby, relevant images should be assigned significantly higher weight values as irrelevant images. The final representation of the input sequence is subsequently computed by performing a weighted aggregation of the BiLSTM output $\{h_{1,j}, h_{2,j}, \dots, h_{l,j}\}$ based on the computed self-attention weights $\{a_k\}_{k=1}^l$ as follows,

$$h_j = \sum_{k=1}^l a_k \cdot h_{k,j} \quad (20)$$

and is further used to perform the classification task.

4. Experiments

In the following section, a description of the experiments performed for the evaluation of the proposed approach is provided. First, the datasets used for the evaluation are briefly described, followed by a depiction of the conducted data preprocessing steps. The experimental settings as well as the performed experiments are described subsequently. This section is finally concluded with a description and discussion of the experimental results.

4.1. Datasets Description

The presented approach is evaluated on both the *BioVid Heat Pain Database (Part A)* (BVDB) [4] and the *SenseEmotion Database* (SEDB) [6]. Both datasets were recorded with the principal goal of fostering research in the domain of pain recognition. In both cases, several healthy participants were submitted to a series of individually calibrated heat-induced painful stimuli, using the exact same procedure. Whereas the BVDB consists of 87 individuals submitted to four individually calibrated and gradually increasing levels of heat-induced painful stimuli (T_1 , T_2 , T_3 and T_4), the SEDB consists of 40 individuals submitted to three individually calibrated and gradually increasing levels of heat-induced stimuli (T_1 , T_2 and T_3). Each single level of heat-induced pain stimulation was randomly elicited a total of 20 times for the BVDB and 30 times for the SEDB. Each elicitation lasted 4 s, followed by a recovery phase of a random length of 8 to 12 s during which a baseline temperature T_0 (32°C) was applied (see Figure 5). Whereas the elicitations were performed uniquely on one specific hand for the BVDB, the experiments were conducted twice for the SEDB, with the elicitation performed each time on one specific arm (left forearm and right forearm). Therefore, with the inclusion of the baseline temperature T_0 , the dataset specific to the BVDB consists of a total of $87 \times 20 \times 5 = 8700$ samples, whereas the SEDB consists of a total of $40 \times 30 \times 4 \times 2 = 9600$ samples. During the experiments, the demeanour of each participant was recorded using several modalities consisting of video and bio-physiological channels

for the BVDB, while the SEDB included audio, video and bio-physiological channels. The current work focuses uniquely on the video modality, and the reader should refer to the work in [10,14–16,33,59–64] for more experiments including the other recorded modalities.

4.2. Data Preprocessing

The evaluation performed in the current work is undertaken in both cases (BVDB and SEDB) on video recordings performed by a frontal camera. The recordings were performed at a frame rate of 25 frames per second (fps) for the BVDB and 30 fps for the SEDB. Furthermore, the evaluation is performed uniquely on windows of length 4.5 s with a shift of 4 s from the elicitation's onset, as proposed in [16] (see Figure 5). Once these specific windows are extracted, the facial behaviour analysis toolkit OpenFace [65] is used for the automatic detection, alignment and normalisation of the facial area (with a fixed size of 100×100 pixels) in each video frame. Subsequently, MHI sequences and OFI sequences are extracted using the OpenCV library [66]. Both MHIs and OFIs are generated relatively to a reference frame, which in this case is the very first frame of each video sequence. Concerning MHIs, the temporal extent parameter τ (see Equation (1)) was set to the length of the sequence of images ($25 \times 4.5 \cong 113$ frames for the BVDB and $30 \times 4.5 = 135$ frames for the SEDB). Furthermore, the threshold parameter ξ (see Equation (2)) was set to 1 to capture any single motion from two consecutive frames (in this case, the fluctuation of pixel intensities between the reference frame and the t th frame). Finally, to reduce the computational requirements, the number of samples in each sequence is reduced by sequentially selecting each second frame of an entire sequence for the BVDB (resulting in sequences with a total length of 57 frames), and each third frame of an entire sequence for the SEDB (resulting in sequences of length 45 frames). The dimensionality of the tensor input specific to the BVDB is, respectively, $(bs, 57, 100, 100, 3)$ for OFI sequences and $(bs, 57, 100, 100, 1)$ for MHI sequences (bs representing the batch size). The dimensionality of the tensor input specific to the SEDB is, respectively, $(bs, 45, 100, 100, 3)$ for OFI sequences and $(bs, 45, 100, 100, 1)$ for MHI sequences.

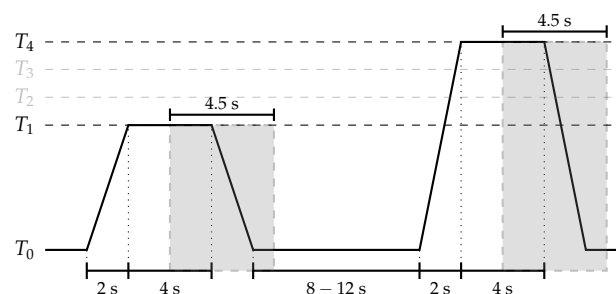


Figure 5. Video Signal Segmentation (BioVid Heat Pain Database (Part A)). Experiments are carried out on windows of length 4.5 s with a temporal shift of 4 s from the elicitations' onsets.

4.3. Experimental Settings

The evaluation performed in the current work consists of the discrimination between high and low stimuli levels. Therefore, two binary classification tasks are performed for each database: T_0 vs. T_4 and T_1 vs. T_4 for the BVDB, and T_0 vs. T_3 and T_1 vs. T_3 for the SEDB. Furthermore, the assessment of the proposed approach is conducted by applying a *Leave-One-Subject-Out* (LOSO) cross-validation evaluation, which means that a total of 87 experiments were conducted for the BVDB (40 experiments for the SEDB), during which the data specific to each participant is used once to evaluate the performance of the classification architecture optimised on the data specific to the remaining 86 participants (the data specific to 39 participants is used to optimise the architecture for the SEDB, and the data specific to the remaining participant is used to evaluate the performance of the architecture).

The feature embedding CNN used for the evaluation is adapted from the one proposed by the Visual Geometry Group of the University of Oxford *VGG16* [67]. The depth of the CNN model is

substantially reduced to a total of 10 convolutional layers (instead of 13 as in the *VGG16* model), and the number of convolutional filters is gradually increased from one convolutional block to the next starting from 8 filters until a maximum of 64 filters. The activation function in each convolutional layer consists of the *elu* activation function (instead of the rectified linear unit (*relu*) activation function as in the *VGG16* model). Max-pooling and Batch Normalisation [68] are performed after each convolutional block. A detailed description of the feature embedding CNN architecture can be seen in Table 1. The coupled BiLSTM layer consists of two LSTM RNNs with 64 units each. The resulting sequence of spatio-temporal features is further fed into the attention layer in order to generate a single aggregated representation of the input sequence. The classification is further performed based on this representation and the architecture of the classification model is described in Table 2. The exact same architecture is used for the two input sequences (MHIs and OFIs). The outputs of the classifiers are further aggregated based on both Equation (8) and Equation (9). The whole architecture is subsequently trained in an end-to-end manner, using the Adaptive Moment Estimation (Adam) [69] optimisation algorithm with a fixed learning rate set empirically to 10^{-5} . The categorical cross entropy loss function is used for each network ($\mathcal{L}_{mhi} = \mathcal{L}_{ofi} = \mathcal{L}_{agg} = \mathcal{L}$), and is defined as follows,

$$\mathcal{L} = - \sum_{j=1}^c y_j \log(\hat{y}_j) \quad (21)$$

where \hat{y}_j represents the classifier's output, y_j is the ground-truth label value and $c \in \mathbb{N}$ is the number of classes for a specific classification task.

Table 1. Feature embedding CNN architecture.

Layer	No. Filters
2 × Conv2D	8
MaxPooling2D	–
Batch Normalisation	–
2 × Conv2D	16
MaxPooling2D	–
Batch Normalisation	–
3 × Conv2D	32
MaxPooling2D	–
Batch Normalisation	–
3 × Conv2D	64
MaxPooling2D	–
Batch Normalisation	–
Flatten	–

The size of the kernels is identical for all convolutional layers and is set to 3×3 , with the convolutional stride set to 1×1 . Max-pooling is performed after each block of convolutional layers over a 2×2 window, with a 2×2 stride.

The regularisation parameters of the loss function in Equation (10) are set as follows: $\lambda_{mhi} = \lambda_{ofi} = 0.2$ and $\lambda_{agg} = 0.6$. The value of the regularisation parameter specific to the aggregation layer's loss is set higher than the others in order to enable the architecture to compute robust aggregation weights. The whole architecture is trained for a total of 20 epoches with the batch size set to 40 for the BVDB and 60 for the SEDB. The implementation and evaluation of the whole architecture is done with the Python libraries Keras [70], Tensorflow [71] and Scikit-learn [72].

Table 2. Classifier Architecture.

Layer	No. Units
Dropout	–
Fully Connected	64
Dropout	–
Fully Connected	c

The dropout rate is empirically set to 0.25. The first fully connected layer uses the *elu* activation function, while the last fully connected layer consists of a *softmax* layer (whereby c depicts the number of classes of the classification task).

4.4. Results

The performance of the classification architectures specific to each input channel (MHIs and OFIs), as well as the performance of the weighted score aggregation approach are depicted in Figure 6. The performance metric in this case is the accuracy, which is defined as

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (22)$$

where tp refers to true positives, tn refers to true negatives, fp refers to false positives and fn refers to false negatives (since we are dealing with a binary classification task with two balanced datasets). For both datasets and both classification tasks, the aggregation approach significantly outperforms the classification architecture based uniquely on MHIs. Furthermore, the classification architecture based uniquely on OFIs outperforms the one based on MHIs for both databases and both classification tasks, with significant performance improvement in the case of the BVDB. The aggregation approach also performs slightly better than the architecture based uniquely on OFIs for both databases, although not significantly in most cases. The only significant performance improvement is achieved for the classification task T_1 vs. T_4 for the SEDB. However, the performance of both channel specific architectures and the performance of the score aggregation approach are significantly higher than chance level (which is 50% in the case of binary classification tasks) pointing at the relevance of the designed approach. Furthermore, the performance of the classification architecture is improved by using both channels and performing a weighted aggregation of the scores of both channel specific deep attention models.

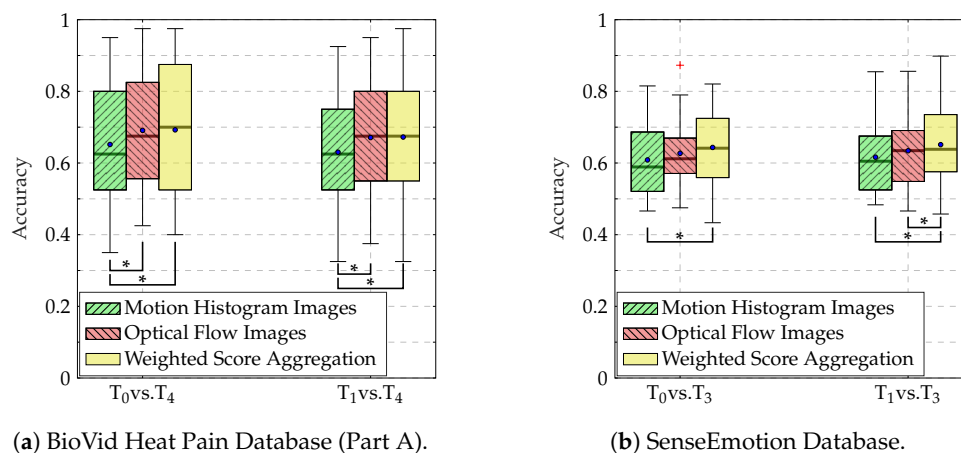


Figure 6. Classification performance (Accuracy). An asterisk (*) indicates a significant performance improvement. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5%. Within each boxplot, the mean and the median classification accuracy are depicted respectively with a dot and a horizontal line.

Moreover, to provide more insights into the self attention mechanism, the frame attention weight values computed at each evaluation step during the LOSO cross-validation evaluation process are depicted in Figure 7 for the BVDB and in Figure 8 for the SEDB (uniquely for the classification task T_0 vs. T_4 , as the results for the classification task T_1 vs. T_4 are similar). The distribution of the weight values specific to the MHI deep attention models for both databases (Figure 7a,c for the BVDB, Figure 8a,c for the SEDB) is skewed left. It depicts a steady growth of weight values along the temporal axis of each sequence, with the MHIs located at the end of a sequence weighted significantly higher as the others. This is in accordance with the sequential extraction process of MHIs, as each extracted image contains more motion information as the previous one, with the last images accumulating almost the totality of motion information of an entire sequence. Therefore, concerning the actual classification task, the last MHIs are more interesting and relevant than the early images. Thus, such images should be weighted accordingly higher. The designed network is therefore capable of conducting this specific task by using self attention mechanisms.

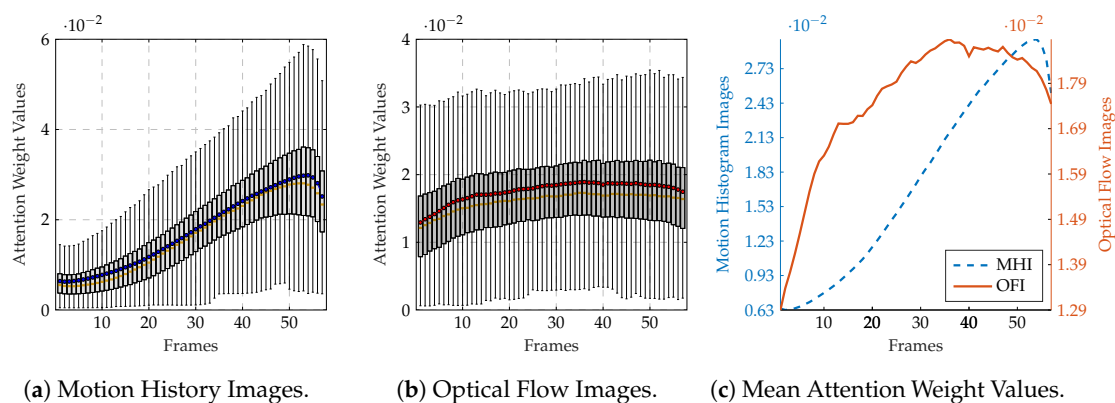


Figure 7. BioVid Heat Pain Database (Part A): Attention network weight values for the classification task T_0 vs. T_4 . Within each boxplot in (a,b), the mean and the median weight values are depicted, respectively, with a dot and a horizontal line. In (c), the average weight values are normalised between the maximum average value and the minimum average value to allow a better visualisation of the values distributions.

A similar observation can be made concerning the distribution of the weight values of OFIs (see Figure 7b,c for the BVDB, Figure 8b,c for the SEDB). Both depicted distributions are also skewed left, with gradually increasing weight values relative to the temporal axis. This shows that the recorded pain-related facial expressions for both BVDB and SEDB consist of gradually evolving facial movements, starting from a neutral facial depiction (not relevant for the actual classification task) to the apex of the facial movement (which is the most relevant frame for the depicted facial emotion) before gradually turning back to the neutral facial depiction. Therefore, the network assigns weight values according to this specific characterisation of pain-related facial movements using attention mechanisms, thus the relevance of such approaches for facial expression analysis.

Furthermore, the performance of the weighted score aggregation approach is further assessed based on the following additional performance metrics,

$$\text{Macro Precision} = \frac{1}{c} \sum_{i=1}^c \frac{tp_i}{tp_i + fp_i} \quad (23)$$

$$\text{Macro Recall} = \frac{1}{c} \sum_{i=1}^c \frac{tp_i}{tp_i + fn_i} \quad (24)$$

$$\text{Macro F1 Score} = \frac{2 \times \text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}} \quad (25)$$

where tp_i , fp_i and fn_i refer, respectively, to the true positives, false positives and false negatives of the i th class. The results of the evaluation are depicted in Figure 9, for both the BVDB (see Figure 9a) and the SEDB (see Figure 9b).

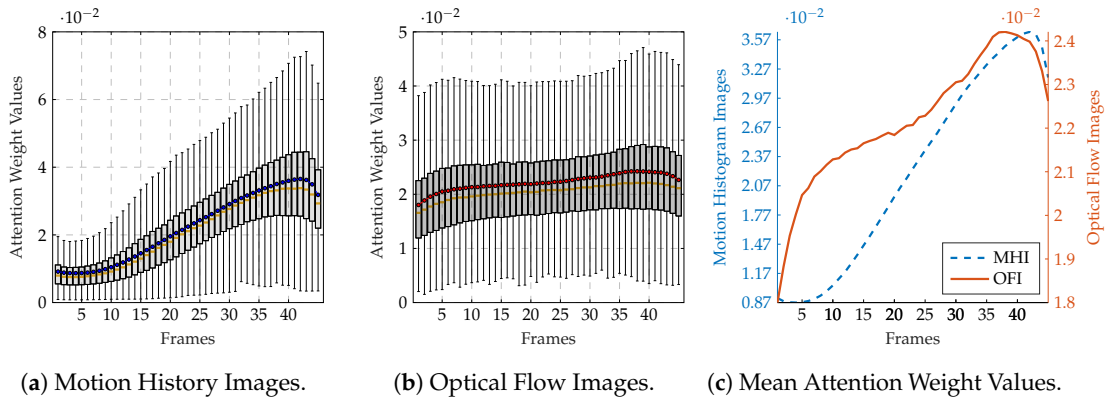


Figure 8. SenseEmotion Database: Attention network weight values for the classification task T_0 vs. T_3 . Within each boxplot in (a,b), the mean and the median weight values are depicted respectively with a dot and a horizontal line. In (c), the average weight values are normalised between the maximum average value and the minimum average value to allow a better visualisation of the values distributions.

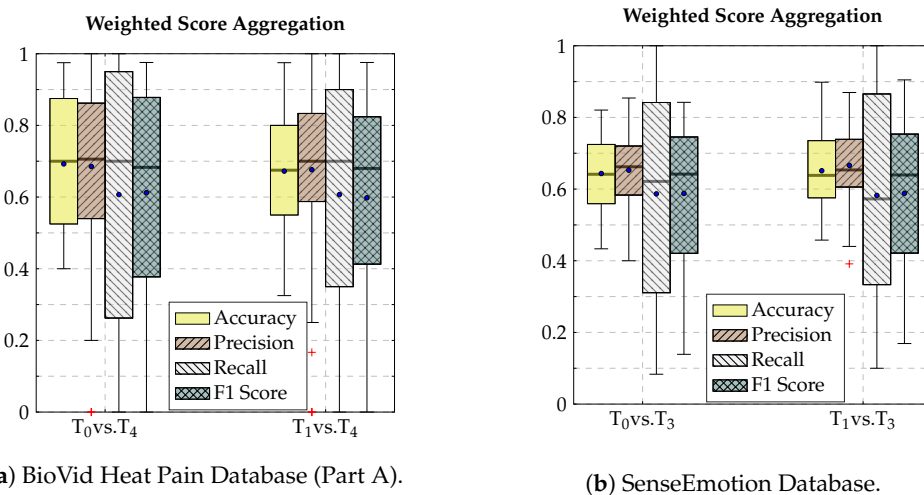


Figure 9. Weighted score aggregation classification performance. Within each box plot, the mean and median values of the respective performance evaluation metrics are depicted with a dot and a horizontal line, respectively.

These results depict a huge variance amongst all performance metrics, in particular the *Macro Recall*, which points at the fact that the classification tasks remain difficult. The evaluation on some participants yields a *Macro F1 Score* of null or nearly null, pointing at the fact that the architecture is unable to discriminate between low and high levels of pain elicitation for these specific participants. This is, however, similar and in accordance with previous works on these specific datasets. The authors of the BVDB in [73] were able to identify a set of participants who did not react to the levels of pain elicitation, therefore causing the huge variance in the classification experiments.

Finally, the performance of the weighted score aggregation approach is compared to other pain-related facial expressions classification approaches proposed in the literature. For the sake of fairness, we compare the results of the proposed approach with those results in related works which are based on the exact same dataset and were computed based on the exact same evaluation protocol (LOSO). The results are depicted in Table 3 for the BVDB and in Table 4 for the SEDB.

Table 3. Classification performance comparison to early works on the BioVid Heat Pain Database (Part A) in a LOSO cross-validation setting for the classification task T_0 vs. T_4 .

Approach	Description	Performance
Yang et al. [27]	BSIF	65.17
Kächele et al. [31,62]	Geometric Features	65.55 ± 14.83
Werner et al. [8]	Standardised Facial Action Descriptors	72.40
Our Approach	Motion History Images	65.17 ± 15.49
Our Approach	Optical Flow Images	69.11 ± 14.73
Our Approach	Weighted Score Aggregation	<u>69.25 ± 17.31</u>

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold and the second best approach is underlined.

Table 4. Classification performance comparison to early works on the SenseEmotion Database in a LOSO cross-validation setting for the classification task T_0 vs. T_3 .

Approach	Description	Performance
Kalischek et al. [38]	Transfer Learning	60.10 ± 00.06
Thiam et al. [15]	Standardised Geometric Features	66.22 ± 14.48
Our Approach	Motion Histogram Images	60.86 ± 09.81
Our Approach	Optical Flow Images	62.70 ± 09.24
Our Approach	Weighted Score Aggregation	<u>64.35 ± 10.40</u>

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold and the second best approach is underlined.

In both cases, the performance of the weighted score aggregation approach is on par with the best performing approaches. However, it has to be mentioned that the authors of the best performing approaches for both the BVDB [8] and the SEDB [15] perform a subject-specific normalisation of the extracted feature representations in order to compensate for the differences in expressiveness amongst the participants. Although this specific preprocessing step has proven to significantly improve the classification performance of the architecture [61], it is not realistic as it requires that the whole testing set is already available beforehand. The normalisation parameters should be learned on the available training material and subsequently applied to the testing material during the inference phase. Nevertheless, the proposed approach based on the weighted aggregation of the scores of both MHI- and OFI-specific deep attention models generalises well and is capable of achieving state-of-the-art classification performances.

5. Conclusions

In the current work, an approach based on a weighted aggregation of the scores of two deep attention networks based, respectively, on MHIs and OFIs has been proposed and evaluated for the analysis of pain-related facial expressions. The assessment performed on both BVDB and SEDB shows that the proposed approach is capable of achieving state-of-the-art classification performances and is on par with the best performing approaches proposed in the literature. Moreover, the visualisation of the weight values stemming from the implemented attention mechanism shows that the network is capable of identifying relevant frames in relation with the current level of pain elicitation depicted by a sequence of images, by assigning significantly higher values to the most relevant images in comparison to the weight values of irrelevant images. Furthermore, as the proposed architecture was trained from scratch in an end-to-end manner, it is believed that transfer learning, in particular, for the feature embedding CNN used to generate the feature representation of each frame, could potentially improve the performance of the whole architecture. Such an analysis was not conducted in the current

work, as the optimisation of the presented approach was not the goal of the conducted experiments, but rather the assessment of such an architecture for the analysis of pain-related facial expressions. Moreover, a multi-stage training strategy could also potentially improve the overall performance of the architecture, as the end-to-end trained approach is likely to suffer from overfitting, in particular, when considering the coupled aggregation layer. The representation of the input sequences should be further investigated as well. Both MHIs and OFIs have the temporal aspect of the sequences integrated into their properties. The performed evaluation has shown that a model based on OFIs significantly outperforms the one based on MHIs in most cases. However, it has also been shown that most of the interesting frames in MHI sequences are located at the very end of the temporal axis of each sequence. Therefore, single MHIs extracted from entire sequences could also be used as input for deep architectures. Overall, the performed experiments show that the discrimination between lower and higher pain elicitation levels remains a difficult endeavour. This is due to the variety of expressiveness amongst the participants. However, personalisation and transfer learning strategies could potentially help improve the performance of inference models applied in this specific area of research.

Author Contributions: Conceptualisation, P.T. and F.S.; Methodology, P.T.; Software, P.T.; Validation, P.T.; Formal Analysis, P.T.; Investigation, P.T. and F.S.; Writing—Original Draft Preparation, P.T.; Writing—Review and Editing, P.T., H.A.K. and F.S.; Visualisation, P.T.; Supervision, H.A.K. and F.S.; Project Administration, H.A.K. and F.S.; Funding Acquisition, H.A.K. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding from the Federal Ministry of Education and Research (BMBF, SenseEmotion) to F.S., (BMBF, e:Med, CONFIRM, ID 01ZX1708C) to H.A.K., and the Ministry of Science and Education Baden-Württemberg (Project ZIV) to H.A.K.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ahad, M.A.R.; Tan, J.K.; Kim, H.; Ishikawa, S. Motion History Image: its variants and applications. *Mach. Vis. Appl.* **2012**, *23*, 255–281. [[CrossRef](#)]
2. Horn, B.K.P.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [[CrossRef](#)]
3. Lucey, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the Face and Gesture, Santa Barbara, CA, USA, 21–25 March 2011; pp. 57–64.
4. Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.; Traue, H.C.; Crawcour, S.; Werner, P.; Al-Hamadi, A.; Andrade, A. The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In Proceedings of the IEEE International Conference on Cybernetics, Lausanne, Switzerland, 13–15 June 2013; pp. 128–131.
5. Aung, M.S.H.; Kaltwang, S.; Romera-Paredes, B.; Martinez, B.; Singh, A.; Cella, M.; Valstar, M.; Meng, H.; Kemp, A.; Shafizadeh, M.; et al. The automatic detection of chronic pain-related expression: requirements, challenges and multimodal dataset. *IEEE Trans. Affect. Comput.* **2016**, *7*, 435–451. [[CrossRef](#)] [[PubMed](#)]
6. Velana, M.; Gruss, S.; Layher, G.; Thiam, P.; Zhang, Y.; Schork, D.; Kessler, V.; Gruss, S.; Neumann, H.; Kim, J.; et al. The SenseEmotion Database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Cancun, Mexico, 4 December 2016; pp. 127–139.
7. Thiam, P.; Kessler, V.; Schwenker, F. Hierarchical combination of video features for personalised pain level recognition. In Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 26–28 April 2017; pp. 465–470.
8. Werner, P.; Al-Hamadi, A.; Limbrecht-Ecklundt, K.; Walter, S.; Gruss, S.; Traue, H.C. Automatic Pain Assessment with Facial Activity Descriptors. *IEEE Trans. Affect. Comput.* **2017**, *8*, 286–299. [[CrossRef](#)]
9. 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 Proceedings of the Interspeech 2016, San-Francisco, CA, USA, 8–12 September 2016; pp. 92–96.

10. Thiam, P.; Schwenker, F. Combining deep and hand-crafted features for audio-based pain intensity classification. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Beijing, China, 20 August 2018; pp. 49–58.
11. Walter, S.; Gruss, S.; Limbrecht-Ecklundt, K.; Traue, H.C.; Werner, P.; Al-Hamadi, A.; Diniz, N.; Silva, G.M.; Andrade, A.O. Automatic pain quantification using autonomic parameters. *Psych. Neurosci.* **2014**, *7*, 363–380. [[CrossRef](#)]
12. Chu, Y.; Zhao, X.; Han, J.; Su, Y. Physiological signal-based method for measurement of pain intensity. *Front. Neurosci.* **2017**, *11*, 279. [[CrossRef](#)]
13. Lopez-Martinez, D.; Picard, R. Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, HI, USA, 18–21 July 2018; pp. 5624–5627.
14. Thiam, P.; Schwenker, F. Multi-modal data fusion for pain intensity assessment and classification. In Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications, Montreal, QC, Canada, 28 November–1 December 2017; pp. 1–6.
15. Thiam, P.; Kessler, V.; Amirian, M.; Bellmann, P.; Layher, G.; Zhang, Y.; Velana, M.; Gruss, S.; Walter, S.; Traue, H.C.; et al. Multi-modal pain intensity recognition based on the SenseEmotion Database. *IEEE Trans. Affect. Comput.* **2019**. [[CrossRef](#)]
16. Thiam, P.; Bellmann, P.; Kestler, H.A.; Schwenker, F. Exploring deep physiological models for nociceptive pain recognition. *Sensors* **2019**, *19*, 4503. [[CrossRef](#)]
17. Ekman, P.; Friesen, W.V. *The Facial Action Unit System: A Technique for the Measurement of Facial Movement*; Consulting Psychologist Press: Mountain View, CA, USA, 1978.
18. Senechal, T.; McDuff, D.; Kaliouby, R.E. Facial Action Unit detection using active learning and an efficient non-linear kernel approximation. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 10–18.
19. Lucey, P.; Cohn, J.; Lucey, S.; Matthews, I.; Sridharan, S.; Prkachin, K.M. Automatically detecting pain using Facial Actions. In Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–8.
20. Abe, S. *Support Vector Machines for Pattern Classification*; Springer: Berlin, Germany, 2005.
21. Brümmer, N.; Preez, J.D. Application-independent evaluation of speaker detection. *Comput. Speech Lang.* **2006**, *20*, 230–275. [[CrossRef](#)]
22. Zafar, Z.; Khan, N.A. Pain intensity evaluation through Facial Action Units. In Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4696–4701.
23. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [[CrossRef](#)]
24. Prkachin, K.M.; Solomom, P.E. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* **2008**, *139*, 267–274. [[CrossRef](#)]
25. Xu, X.; Craig, K.D.; Diaz, D.; Goodwin, M.S.; Akcakaya, M.; Susam, B.T.; Huang, J.S.; de Sa, V.S. Automated pain detection in facial videos of children using human-assisted transfer learning. In Proceedings of the International Workshop on Artificial Intelligence in Health, Stockholm, Sweden, 13–14 July 2018; pp. 162–180.
26. Monwar, M.; Rezaei, S. Pain recognition using artificial neural network. In Proceedings of the IEEE International Symposium on Signal Processing and Information Theory, Vancouver, BC, Canada, 27–30 August 2006; pp. 8–33.
27. Yang, R.; Tong, S.; Bordallo, M.; Boutellaa, E.; Peng, J.; Feng, X.; Hadid, A. On pain assessment from facial videos using spatio-temporal local descriptors. In Proceedings of the 6th International Conference on Image Processing Theory, Tools and Applications, Oulu, Finland, 12–15 December 2016; pp. 1–6.
28. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [[CrossRef](#)]
29. Ojansivu, V.; Heikkilä, J. Blur insensitive texture classification using local phase quantization. In Proceedings of the Image and Signal Processing, Cherbourg-Octeville, France, 1–3 July 2008; pp. 236–243.
30. Kannala, J.; Rahtu, E. BSIF: Binarized Statistical Image Features. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012; pp. 1363–1366.

31. Kächele, M.; Thiam, P.; Amirian, M.; Werner, P.; Walter, S.; Schwenker, F.; Palm, G. Engineering Applications of Neural Networks. Multimodal data fusion for person-independent, continuous estimation of pain Intensity. In Proceedings of the Engineering Applications of Neural Networks, Rhodes, Greece, 25–28 September 2015; pp. 275–285.
32. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
33. Thiam, P.; Kessler, V.; Walter, S.; Palm, G.; Scwenker, F. Audio-visual recognition of pain intensity. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Cancun, Mexico, 4 December 2016; pp. 110–126.
34. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.
35. Almaev, T.R.; Valstar, M.F. Local Gabor Binary Patterns from Three Orthogonal Planes for automatic facial expression recognition. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 356–361.
36. Bellantonio, M.; Haque, M.A.; Rodriguez, P.; Nasrollahi, K.; Telve, T.; Guerrero, S.E.; González, J.; Moeslund, T.B.; Rasti, P.; Anbarjafari, G. Spatio-temporal pain recognition in CNN-based super-resolved facial images. In Proceedings of the International Conference on Pattern Recognition: Workshop on Face and Facial Expression Recognition, Cancun, Mexico, 4 December 2016; pp. 151–162.
37. Rodriguez, P.; Cucurull, G.; González, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, F.X. Deep Pain: Exploiting Long Short-Term Memory networks for facial expression classification. *IEEE Trans. Cybern.* **2018**. [[CrossRef](#)]
38. Kalischek, N.; Thiam, P.; Bellmann, P.; Schwenker, F. Deep domain adaptation for facial expression analysis. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Cambridge, UK, 3–6 September 2019; pp. 317–323.
39. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and application in vision. In Proceedings of the IEEE International Symposium on Circuits and Systems, 2010, Paris, France, 30 May–2 June 2010; pp. 253–256.
40. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
41. Soar, J.; Bargshady, G.; Zhou, X.; Whittaker, F. Deep learning model for detection of pain intensity from facial expression. In Proceedings of the International Conference on Smart Homes and Health Telematics, Singapore, 10–12 July 2018; pp. 249–254.
42. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, Williams College, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
43. Bargshady, G.; Soar, J.; Zhou, X.; Deo, R.C.; Whittaker, F.; Wang, H. A joint deep neural network model for pain recognition from face. In Proceedings of the IEEE 4th International Conference on Computer and Communication Systems, Singapore, 23–25 February 2019; pp. 52–56.
44. Zhou, J.; Hong, X.; Su, F.; Zhao, G. Recurrent convolutional neural network regression for continuous pain intensity estimation in Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1535–1543.
45. Liang, M.; Hi, X. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.
46. Wang, F.; Xiang, X.; Liu, C.; Tran, T.D.; Reiter, A.; Hager, G.D.; Quaon, H.; Cheng, J.; Yuille, A.L. Regularizing face verification nets for pain intensity regression. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 1087–1091.
47. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame attention networks for facial expression recognition in videos. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870.
48. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [[CrossRef](#)]
49. Yin, Z.; Collins, R. Moving object localization in thermal imagery by forward-backward MHI. In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, New York, NY, USA, 17–22 June 2006; pp. 133–140.

50. Farneback, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; pp. 363–370.
51. Brox, T.; Bruhn, A.; Papenber, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 25–36.
52. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, University of British Columbia, Vancouver, BC, Canada, 24–28 August 1981; pp. 674–679.
53. Beauchemin, S.S.; Barron, J.L. The computation of optical flow. *ACM Comput. Surv.* **1995**, *27*, 433–466. [[CrossRef](#)]
54. Akpinar, S.; Alpaslan, F.N. Chapter 21—Optical flow-based representation for video action detection. In *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*; Deligiannidis, L., Arabnia, H.R., Eds.; Morgan Kaufmann: Boston, MA, USA, 2015; pp. 331–351.
55. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038. [[CrossRef](#)]
56. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Network. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [[CrossRef](#)]
57. Hochreiter, S.; Bengio, Y.; Frasconi, P. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *Field Guide to Dynamical Recurrent Networks*; IEEE Press: Piscataway, NJ, USA, 2001.
58. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2016**, arXiv:1511.07289. Available online: <https://arxiv.org/abs/1511.07289> (accessed on 3 February 2020) [[CrossRef](#)]
59. Werner, P.; Al-Hamadi, A.; Niese, R.; Walter, S.; Gruss, S.; Traue, H.C. Automatic pain recognition from video and biomedical signals. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4582–4587.
60. Walter, S.; Gruss, S.; Traue, H.; Werner, P.; Al-Hamadi, A.; Kächele, M.; Schwenker, F.; Andrade, A.; Moreira, G. Data fusion for automated pain recognition. In Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, Istanbul, Turkey, 20–23 May 2015; pp. 261–264.
61. Kächele, M.; Thiam, P.; Amirian, M.; Schwenker, F.; Palm, G. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE J. Sel. Top. Sign. Process.* **2016**, *10*, 854–864. [[CrossRef](#)]
62. Kächele, M.; Amirian, M.; Thiam, P.; Werner, P.; Walter, S.; Palm, G.; Schwenker, F. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evol. Syst.* **2016**, *8*, 1–13. [[CrossRef](#)]
63. Bellmann, P.; Thiam, P.; Schwenker, F. Computational Intelligence for Pattern Recognition. In *Computational Intelligence for Pattern Recognition*; Pedrycz, W., Chen, S.M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 83–113.
64. Bellmann, P.; Thiam, P.; Schwenker, F. Using a quartile-based data transform for pain intensity classification based on the SenseEmotion Database. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Cambridge, UK, 3–6 September 2019; pp. 310–316.
65. Baltrusaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
66. Bradski, G. The OpenCV library. *Dr Dobb's J. Softw. Tools* **2000**, *25*, 120–125.
67. Simonyan, K.; Zisserman, A. Very deep convolution networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556. Available online: <https://arxiv.org/abs/1409.1556> (accessed on 3 February 2020) [[CrossRef](#)]
68. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167. Available online: <https://arxiv.org/abs/1502.03167> (accessed on 3 February 2020) [[CrossRef](#)]
69. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980. Available online: <https://arxiv.org/abs/1412.6980> (accessed on 3 February 2020) [[CrossRef](#)]
70. Chollet, F. Keras. 2015. Available online: <https://keras.io> (accessed on 21 January 2020).

71. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, C.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: <https://www.tensorflow.org/> (accessed on 21 January 2020).
72. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
73. Werner, P.; Al-HamadiAl-Hamadi, A.S. Analysis of facial expressiveness during experimentally induced heat pain. In Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, San Antonio, TX, USA, 23–26 October 2017; pp. 176–180.

Sample Availability: The BioVid Heat Pain Database (Part A) is publicly available for non-commercial research and can be acquired by contacting the authors of the database at the following web-page: <http://www.iikt.ovgu.de/BioVid.print>.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).