Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification

Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M. Gonfaus, Kamal Nasrollahi, Thomas B. Moeslund, F. Xavier Roca

Abstract—Pain is an unpleasant feeling that has been shown to be an important factor for the recovery of patients. Since this is costly in human resources and difficult to do objectively, there is the need for automatic systems to measure it. In this paper, contrary to current state-of-the-art techniques in pain assessment, which are based on facial features only, we suggest that the performance can be enhanced by feeding the raw frames to deep learning models, outperforming the latest state-of-the-art results while also directly facing the problem of imbalanced data. As a baseline, our approach first uses convolutional neural networks (CNN) to learn facial features from VGG_Faces, which are then linked to a Long Short-Term Memory (LSTM) to exploit the temporal relation between video frames. We further compare the performances of using the so popular schema based on the canonically normalized appearance versus taking into account the whole image: As a result, we outperform current state-of-the-art AU performance in the UNBC-McMaster Shoulder Pain Expression Archive Database. In addition, to evaluate the generalization properties of our proposed methodology on facial motion recognition, we also report competitive results in the Cohn Kanade+ facial expression database.

I. INTRODUCTION

The automatic detection of pain is a subject of high interest in the health domain since it is not only an important indicator for medical diagnosis, but has also been shown to be an obstacle for patient recuperation in Intensive Care Units [1] and after surgery [2]. In [3], it is shown how good pain assessment is crucial for a good pain control, which is usually verbally checked by professional nurses, known as self-report. However, this is not always possible due to the age of the patient, the particular illness or language impairments. Moreover, pain is a subjective feeling which can be described differently across cultures [4]. Thus, pain assessment could be highly benefited from automatic tools.

Indeed this goal has been already addressed several times in the past, for example in 2011 the authors of [5] tackle the problem using brain activity imaging. So pain detection is also an important task from the point of view of computer vision, since it is a clear step towards an automatic detector of spontaneous face expressions [6], [7], [8], [9] and [10]. In particular, it was of high importance for the computer vision community the release of a database published by Lucey et al. in [11], in order to alleviate the lack of representative data of the other existing databases. Their UNBC-McMaster database consists of 200 video sequences taken from 25 patients who were suffering from shoulder pain. The frames were labeled using the validated Prkachin and Solomon metric [12] (PSPI) based on the Facial Action Coding System (FACS) [13], which codes different movements of the face muscles with different intensity levels. It is a very challenging dataset, and as it can be seen in Fig. 1, in some cases it can be very hard to determine whether a subject is in pain or not, even for clinical professionals.

So the UNBC-McMaster Painful dataset has been used to propose new models for facial pain detection. In the first place, Lucey et al. in [11] already released a baseline along with the dataset, using Support Vector Machines (SVMs) on top of the pixel and landmark features extracted using Active Appearance Models (AAM) [14] in order to predict painful Action Units (AUs) and the PSPI for the presence of pain. [10] proposed a late fusion of shape and appearance features in order to predict the continuous PSPI scores of the Painful data.

In fact, facial Action Units have been typically used to encode facial motion corresponding to different facial expressions such as pain or anger. As stated by Rudovic et al. [15], the task of AU intensity estimation is very challenging, due to the high variability in facial expressions depending on the context, such as illumination, head movements or subject-specific expressions. Being a complex task, Action Unit intensity estimation has received a lot of attention over two decades for generic facial motion analysis. It has been approached by Kim et al. in [16], where they use a dynamic ranking model to overcome the difficulty of the emotion intensities differing substantially across subjects. Valstar et al. [17] also tackle the task of facial AUs recognition by using a facial point detector to localize 20 facial fiducial points. Then these points are tracked through a sequence of images and then a combination of GentleBoost, SVMs and hidden Markov models (HMM) is used for AU recognition. According to [18] most of the temporal graphical models such as HMM or conditional random fields (CRF) used for AU recognition fail to jointly model different emotions. To overcome this issue, they propose the use of a Hidden Conditional Ordinal Random Field (H-CORF) to achieve both intensity estimation of facial expressions and dynamic recognition of multiple emotions at the same time. Ming et al. [19] proposed a method based
on multi-kernel SVM and feature fusion to approach AUs intensity estimation.

Focused on facial landmark estimation for pain detection, Rudovic et al. [20] proposed to use a heteroscedastic Conditional Ordinal Random Field (CORF) model in order to deal with the inter-subject variability of the pain expressions. Authors in [21] and [22] used weakly supervised learning and multiple instance learning to predict pain only using sequence-level annotations. Khan et al. [23] also used the referenced dataset for pain/no-pain recognition using shape information extracted with a pyramid histogram of orientation gradients (PROG) and appearance information using a pyramid local binary patterns (PLBP). Subsequently, Zafar and Khan’s [24] used a K-NN classifier to classify AUs using 22 facial characteristic points. In 2015, Irani et al. [25] use Spatiotemporal Feature Extraction in order to model the exploits the released energy of the facial muscles in the spatial and temporal domains. They applied their system to both RGB [25] and RGB-Thermal-Depth [26] facial images. Presti and Cascia [27] use Hankel Matrices to represent the temporal dynamics of a sequence of Face Image Descriptors. Pedersen [28] addressed the identity bias of the dataset using autoencoders, ensuring the presence of discriminative features by training with a combined loss function that balances the reconstruction error and the classification error. Later in the same year, Neshov and Manolova [29] used SVMs on top of Scale Invariant Feature Transform (SIFT) features for continuous and discrete PSPI prediction. Rathee and Ganotra [30] proposed the use of Thin Plate Spline (TPS) mapping [31] for modeling the deformation of facial features and a Distance Metric Learning (DML) method to ensure the distance between features belonging to different levels of pain. The recent work of Zhao et al. [32] proposes the novel Alternating Direction Method of Multipliers (ADMM) to solve Ordinal Support Vector Regression (OSVR) achieving competitive performances in supervised, semi-supervised and unsupervised prediction of PSPI scores.

In this paper we continue current trends on deep learning applied to pain estimation [33][34][35][36][37] applied to pain estimation [38]. Similarly to [38], we also perform regression with Deep Convolutional Neural Networks (DCNNs) in order to predict the PSPI score for each frame. Subsequently, we adapt the resulting CNN model for pain classification inspired by [11]. In order to alleviate the problem of data scarcity, we use VGG_Faces, i.e. a VGG-16 CNN [37] pre-trained with millions of faces [39], which already obtains state-of-the-art scores compared with other leave-one-subject-out methods.

Differently to [38], we follow the ideas exposed in [25], by directly exploiting the temporal axis information using Long Short-Term Memory (LSTM) [40][41] on top of the previously-learned deep features, boosting our scores even more. So the main difference of our Deep Learning methodology as described above and the Recurrent Convolutional Neural Networks used in [38] is that we leverage the temporal information without renouncing to the representational power of generic pre-trained CNN features like the ones learned from VGG_Faces, i.e. we link the VGG_Faces features to the LSTM Recurrent Network. In other words, the approach of [38] either discards the temporal information of the data when considering pre-trained features from VGG_Faces or considers temporal information but using less-discriminative features, since the RCNN is learned from scratch.

In addition, differently to [38], we consider the raw image as the input of the CNN, rather than using facial landmarks. By doing so, the proposed method is able to outperform current state-of-the-art in pain intensity estimation.

As pain detection is a form of facial expression recognition, similar methods can be applied to the more general task of emotion recognition. For example Lucey et al. [42] used an SVM on top of features extracted using AAM to build a facial emotion classifier. Based on the observation that only a few facial patches are important for expression recognition, Zhong et al. [43] use a two-stage approach. First LBP features are
used to describe every patch on a grid of $8 \times 8$ over the images of $96 \times 96$ pixels. Then Multi-task sparse learning (MTSL) is used to learn common patches across expressions. Similar to this idea, Liu et al. [44] propose a method which adapts 3D Convolutional Neural Networks (3D CNN) to detect facial action parts under spatial constraints. In the work by Liu et al. [45] they propose to use a Boosted Deep Belief Network to jointly learn the best set of features to describe expression related facial appearance and a classifier on top of these features to perform emotion recognition. Jung et al. [32] also used a deep learning approach, but in this case, they use only one CNN, with the difference that it has several Inception modules. In the work by Zhao et al. [32] they propose the Peak-Piloted Deep Network (PPDN) to use the peak samples (frames with maximum expression) to supervise the feature responses for the non-peak frames of the same emotion and the same subject. Their approach is to supervise the feature responses for the non-peak frames of the representations of non-peak frames.

Contrary to most of the approaches and in the same line as Kaltwang et al. [10], we try to solve the regression task because it fits best to this problem. However, we finally threshold the predictions in order to get performance metrics so that we can compare to [47] or [11] as previously seen in the introduction. The following sub-sections go through the steps of the system.

- **Data Pre-processing.** As it can be seen in Figures 3 and 4, we use the provided landmarks in order to crop and frontalize the faces. Following the procedure in [11], we use Generalized Procrustes Analysis (GPA) to align the landmarks [48]. This method is no more than an extension of the Procrustes Analysis for comparing more than two ordered sets of landmarks. For the simple case, in order to align two sets $X = \{x_1, x_2, ..., x_n\}$, $Y = \{y_1, y_2, ..., y_n\}$ of $N$ landmarks, one has to (i) move their centroids $\bar{x}, \bar{y}$ to the origin (ii) find their scaling factor $s$:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 + (y_i - \bar{y})^2}{N}} \forall x_i, y_i \in X, Y,$$

(1)

so that we can remove it from the landmarks by dividing them by $s$. Then, one can find the rotation $\theta$ between two sets of landmarks by optimizing the rotation angle needed to minimize the mean squared distance between the two sets. This leads to the following equation:

$$\theta = \tan^{-1} \left( \frac{\sum_{i=1}^{N} (w_i y_i - z_i x_i)}{\sum_{i=1}^{N} (w_i y_i + z_i x_i)} \right).$$

(2)

Then, for $K$ sets of points, the GPA consists in choosing one of the sets as a reference in order to align the rest, use the mean of the alignment as a new reference and repeat the process until the Procrustes distance $d = \sqrt{\sum (x_i - y_i)^2}$ between the new reference and the previous one are below a threshold. Once the final reference is obtained, the images are aligned so that their respective landmarks are aligned to it. Then, Delaunay

![Diagram](image)

**Figure 2: Proposed framework.** Schematic depicting the different stages of our proposed pain detection model.
target triangulation is used to create a mesh corresponding to the dual graph of the Voronoi diagram of the points so that piecewise-affine warping can be used to get the so-called canonical normalized appearance. As it can be seen in Fig. 4, we did not use all the provided landmarks since it forces too much the facial expression, i.e., eliminates mouth gestures and closed eyes, and we did not want to lose any pain-related information.

Contrary to the procedure described in [11] and followed by others, e.g., [28], we do not grayscale the image and we warp it to $224 \times 244$ because it is the common input size for most deep neural network models after cropping. We do not crop patches during training due to the fact that faces are already aligned so there is no need for translation invariance.

Finally, per-pixel mean subtraction is performed in order to pass real zeros for the black areas to the neural network. Global contrast normalization is then applied to ease the training of the model.

- **Facing imbalanced data** Since there are about 8K pain frames and about 40K labeled as no pain in PSP1 score, it is probable that any model gets biased towards the prediction of no-pain at the cost of missing pain frames. There are two common approaches to overcome this problem: (i) balancing data, (ii) using weighted loss functions. In this work we balance the training data (i) and validate the original validation data, but we also complement the results by giving normalized scores, as proposed by [49] (i.e., balancing the validation dataset). To balance the data, we randomly under-sample the majority class, i.e., the no-pain class, so that both pain and no-pain categories have the same probability to be randomly picked by the training algorithm. To create the training sequences for the RNN we also need to balance the data, but instead of balancing at the frame level, we balance at the sequence level so that there are no frame skips. This means that we sort the frames in time, split them in sequences, and discard entire sequences with no pain in all their frames until they match the number of sequences with pain inside.

- **Target preprocessing** Because MSE is very sensitive and most suited for the cases where Gaussian noise is present, it is good practice to standardize the labels, i.e., the pain levels, before training.

After data is pre-processed, it is used to train a CNN to perform the pain level recognition task. This is achieved by fine-tuning a VGG-16 CNN pre-trained with Faces [39]. Instead of using the log-likelihood objective function, we used $\text{MSE}$ between the predicted label $Y \hat{\text{Y}}$ and the actual label $Y$ in an attempt to make the model get a better insight on pain detection since it is not binary and it actually proved to perform slightly better:

$$ E = \frac{1}{2N} \sum_{n=1}^{N} \| \hat{y}_n - y_n \|_2^2. \quad (3) $$

In order to improve the model generalization, data augmentation is used. This is done by (i) flipping images with 50% probability, and (ii), adding random noise to the reference landmarks before performing piece-wise affine warping in order to introduce small deformations to faces, see Fig. 3.

The masking and the frontalization performed during the pre-process alter the original face, resulting in an image considerably different from a non-processed face like the ones that the CNN used has been pre-trained with. These differences between the pre-training data and the fine-tuning data could affect the results obtained, because the network has learned to extract specific features from raw face images, and it may not be able to extract them from the processed faces. Thus, we also provide results with a network trained with raw faces, similar to the ones used during pre-training, and each frame is processed only to extract a crop around the face, see Fig. 3, and then the mean pixel value is subtracted to each image.

**A. Convolutional Neural Networks**

Convolutional Neural Networks (CNNs) are an architecture of neural networks proposed by LeCun et al., that localize features in images to extract information of the visual content [50]. CNNs are made of different types of layers, stacked on top of each other. The basic layer of a CNN is the convolution layer, which convolves a given tensor of size $W \times H \times D$, with $K$ different filters of size $F \times F \times D$, with a stride of $S$ between convolutions and padding the input with $P$ zeros. This convolution of the input by $K$ filters outputs a tensor with dimensions:

$$ W' \times H' \times D', $$

where

$$ W' = (W - F + 2P) / S + 1, $$

$$ H' = (H - F + 2P) / S + 1, $$

$$ D' = K. $$

The values of the convolution filters are learned by initializing them randomly and updating them by performing gradient descent using the backpropagation algorithm [51]. To compute the error for a given input to the network, the last layer of the network is a loss layer which computes the error between the ground truth label of an input image and the predicted output for that image. This error at the output is backpropagated to previous layers in order to compute the gradients for the weights of previous layers.

This architecture is specially designed to capture 2D information, so it performs very well on images, where pixels intensities are related to their neighbors. The recent increase in computational power provided by GPUs and the availability of large datasets like Imagenet [52] have made the initial CNN implementations evolve to very deep networks [36], [37]. These deep networks have been proven to perform very well in a variety of computer vision tasks such as human action recognition [53], handwritten digit recognition [54] or automatic face detection [55].
LSTM differs from standard RNN because it has a cell state because the gradients tend to either explode or vanish [57]. In practice, it is difficult to train them for long dependencies, but in theory, it is capable of learning long-term dependencies present on sequential data. The back-propagation algorithm [51], called back-propagation through time BPTT [56], is used to train this kind of networks. Since they have to be unrolled, the training of recurrent neural networks is especially suited for sequential data since their neurons do not only have connections (weights) between the next layers but to themselves, which are used to keep information from previous inputs. Since they have to be unrolled, the training of this kind of networks is done with an extension of the backpropagation algorithm [51], called back-propagation through time BPTT [56].

In this work we use LSTM, a type of RNN which is capable of learning long-term dependencies present on sequential data. Standard RNNs are theoretically capable of learning long-term dependencies, but in practice, it is difficult to train them because the gradients tend to either explode or vanish [57]. LSTM differs from standard RNN because it has a cell state controlled by 3 gates, which decide how much information should be let through. These gates are known as forget, input and output gates, see 2. The amount of information that is let through each gate is controlled by a point-wise multiplication and sigmoid function, as the sigmoid function output is between 0 and 1, indicating how much of the information should let through the gate.

At each time-step, the input gate is computed depending on the input to the LSTM for that time-step and the previously hidden state. The cell state candidate is also computed by:

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

$$\hat{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c).$$

Then output of the forget get is computed as:

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

And when the forget and input gates have determined how much information of the previous cell state $C_{t-1}$ and the new cell state candidate $\hat{C}_t$ should be let through, the cell state for the current time-step is computed:

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t.$$ 

Then, the state can be used in order to predict the output of the cell:

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

$$h_t = o_t \odot \tanh(C_t).$$

In order to train the RNN for pain detection, we used the MSE loss since it better suits the nature of the problem, where pain levels have distances in the output space. In case we need to compare in terms of binary accuracy, we can just use a binary threshold. In fact, we empirically found that using the cross-entropy error for binary classification yielded worse performance than just using a threshold after regression. Concretely, we could only reach 81% of accuracy on the test set with the initial settings shown in Table II, which presents a 83.1% for the same model after regression and thresholding.

To train the LSTM, first a feature vector has to be extracted for each image, being this vector the input to the LSTM. We can think of this feature vector as a low-dimensional representation of the image in the feature space. To create this vector for each frame, the frame is processed through the VGG-16 network.
CNN fine-tuned to perform pain level detection and the outputs of the fully-connected layers are used as the encoding for that frame. As it can be seen in Table III, we found that the outputs in the $fc6$ layer had less temporal invariability than the ones from the $fc7$ and thus, the former yielded better performance when fed to the LSTM. Hence, the $fc6$ is always used for comparison with the state-of-the-art. This process results in $M$ feature vectors $v$ where $M$ is the number of frames and $v \in \mathbb{R}^{4096}$ since the $fc6$ layer of the VGG-16 network has 4096 units. Then, the $M$ feature vectors have to be grouped together in sequences of length $\rho$. The sequences are created so that each frame is the last of a sequence once, e.g. if the first sequence is $s_0 = \{v_0, v_1, ..., v_{n-1}, v_n\}$, the next sequence is $s_1 = \{v_1, v_2, ..., v_n, v_{n+1}\}$. Each sequence $s$ is labeled with one label $t$, corresponding to the label of the last frame of the sequence. In the classification task, $t$ is a binary one-hot vector $t \in \{0, 1\}^2$, and for the regression task $t$ is a real number $t \in \mathbb{R}$. As each sequence has only one label, only the hidden state of the last time-step $h_n$ is used to compute the output of the network.

Hence, the label of a frame is predicted taking into account the past $\rho$ frames. For this problem, we found that $\rho = 16$ worked well, and an LSTM [40] RNN is used in order to avoid the problem of gradient vanishing for long sequences. The network is optimized with ADAM since it has proved to be more stable than SGD with momentum [62].

### III. Experiments and Results

As said in the previous sections, we center our experimentation on The UNBC-McMaster Shoulder Pain Expression Archive Database [11]. In addition, we prove the generality of our model by testing it on the Cohn Kanade+ face emotion detection dataset [42] and obtaining competitive results.

#### A. Results on Pain Recognition

A quick skim through the pain detection literature concerning the database will show the reader that there are multiple benchmark procedures. While the original paper [11] and some posterior ones [28] use leave-one-subject-out cross-validation, others like [20], [23], and [30] use k-fold cross-validation or even leave-one-frame-out cross-validation. In addition, Jeni et al. face the problem of data imbalance in [49], proposing normalized metrics that take the skew into account.

In Table I there is a summary of previous approaches to performing pain detection on the same dataset, indicating the method used to extract features and the classifier or regressor trained with those features. It also shows the metric used to evaluate their approach, along with the score obtained and the performance measure. The main difference between most of the listed previous approaches and our approach is that they manually extract a set of features, and then train a model with them, whereas we use an end-to-end deep learning model which learns to extract features from the data and how to combine them to give the correct output. Our approach is also based on Convolution Neural Networks as in [38], but in contrast, we apply temporal modeling using LSTM onto the features learned from the VGG_Faces network. This is different from the method proposed in [38], which discards the temporal information of the data when considering pre-trained features from VGG_Faces, and considers temporal information on low-discriminative features, since the RCNN is learnt from scratch in an unbalanced, quite small dataset (even smaller in [38], since no data augmentation pre-processing is applied).

In this work, we compare within the dataset authors’ scheme: AUC score on leave-one-subject-out cross-validation, since subject-exclusiveness increases the confidence that the model will behave similarly with new data. In addition to comparing our model in a binary setting by using the AUC score, we also test it against other state-of-the-art continuous prediction models with the Intraclass Correlation Coefficient (ICC), Pearson Correlation Coefficient (PCC), the Mean Square Error (MSE), and the Mean Absolute Error (MAE). For the continuous setting, we aggregated the pain levels as indicated in [32] so that the levels 4 and 5 are merged, as well as 6+, that become the 5th level.

#### Table 1: Summary of previous approaches

This table compares the experimental setup of previous approaches to solve the task of automatic pain detection. We compare our method against the previous approaches that have used a subject-exclusive leave-one-subject-out performance measure and do not discard any painful image.

<table>
<thead>
<tr>
<th>Feature descriptors</th>
<th>Classifier</th>
<th>Performance Measure</th>
<th>Metric</th>
<th>Score</th>
<th>Use all images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucey et al. [58]</td>
<td>PTS, APP,</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>AUC</td>
<td>78%</td>
</tr>
<tr>
<td>Lucey et al. [59]</td>
<td>PTS, APP,</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>AUC</td>
<td>78.4%</td>
</tr>
<tr>
<td>Lucey et al. [11]</td>
<td>SPTS, CAPP</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>AUC</td>
<td>83.9%</td>
</tr>
<tr>
<td>Lucey et al. [47]</td>
<td>SPTS, SAPP, CAPP</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>AUC</td>
<td>84.7</td>
</tr>
<tr>
<td>Kaltwang et al. [10]</td>
<td>PTS, DCT, LBP</td>
<td>RVR</td>
<td>Leave one subject out</td>
<td>MSE, PCC, ICC</td>
<td>1.39, 0.59, 0.50</td>
</tr>
<tr>
<td>Florea et al. [60]</td>
<td>HoT</td>
<td>SVR</td>
<td>Leave one subject out</td>
<td>MSE, PCC</td>
<td>1.21, 0.53</td>
</tr>
<tr>
<td>Zhou et al. [38]</td>
<td>learnt</td>
<td>RCNN</td>
<td>Leave one subject out</td>
<td>MSE, PCC</td>
<td>1.54, 0.65</td>
</tr>
<tr>
<td>Zhao et al. [32]</td>
<td>LBP,Gabor</td>
<td>OSVR-L1, OSVR-L2</td>
<td>Leave one subject out</td>
<td>MAE, ICC</td>
<td>0.81, 0.60, 0.56</td>
</tr>
<tr>
<td>Ashraf et al. [9]</td>
<td>S-PTS, S-APP, C-APP</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>Hit rate</td>
<td>82%</td>
</tr>
<tr>
<td>Hammad et al. [61]</td>
<td>CAPP</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>Recall, F1</td>
<td>61%, 57%</td>
</tr>
<tr>
<td>Rudovic et al. [20]</td>
<td>LBP</td>
<td>KCORF</td>
<td>Custom split</td>
<td>F1</td>
<td>40.2%</td>
</tr>
<tr>
<td>Khan et al. [23]</td>
<td>PHOG, PLBP</td>
<td>SVM, DT</td>
<td>10 fold CV</td>
<td>Accuracy</td>
<td>96.4%</td>
</tr>
<tr>
<td>Rathee et al. [30]</td>
<td>TPS</td>
<td>SVM</td>
<td>Leave one frame out</td>
<td>Accuracy</td>
<td>96.0%</td>
</tr>
<tr>
<td>Pedersen et al. [28]</td>
<td>Custom features</td>
<td>SVM</td>
<td>Leave one subject out</td>
<td>AUC, Accuracy</td>
<td>96.5, 86.1%</td>
</tr>
</tbody>
</table>

Most of our approaches are based on Convolution Neural Networks as in [38], but in contrast, we apply temporal modeling using LSTM onto the features learned from the VGG_Faces network. This is different from the method proposed in [38], which discards the temporal information of the data when considering pre-trained features from VGG_Faces, and considers temporal information on low-discriminative features, since the RCNN is learnt from scratch in an unbalanced, quite small dataset (even smaller in [38], since no data augmentation pre-processing is applied).
Table II: Unbalanced and normalized scores. This table reports the accuracy and area under the ROC curve obtained by different versions of our method.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Normalized [49]</th>
<th>Unbalanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Align</td>
<td>77.1 83.2</td>
<td>83.1 83.1</td>
</tr>
<tr>
<td>Align + Fron.</td>
<td>83.2 82.4</td>
<td>86.4 82.1</td>
</tr>
<tr>
<td>Align + Fron. + Data aug.</td>
<td>85.9 89.9</td>
<td>88.8 89.9</td>
</tr>
<tr>
<td>Aligned Crop</td>
<td>80.8 90.0</td>
<td>87.5 89.6</td>
</tr>
<tr>
<td>Aligned Crop + LSTM</td>
<td>83.8 90.1</td>
<td>90.3 91.3</td>
</tr>
</tbody>
</table>

Figure 5: Average saliency map and average face. The first picture (a) corresponds to the average saliency map computed for each image as described by Simonyan et al. [63], the second picture is the average of all the training images. The saliency map shows where the CNN is looking to decide the level of pain of a frame.

In our case, and only for comparison purposes, we also trained on aligned and canonical normalized faces but including data augmentation to add robustness to the model predictions. In Table II, we show the effect of the different stages of pre-processing shown in Fig. 3 on the performance of the model. Specifically, it can be seen that the aligned frontalized facial landmarks proposed in [11] already provides a good performance, but the VGG_faces model is not pre-trained with similar kind of images [39]. In fact, it is interesting that with canonically normalized appearance, the position and translation invariances of the faces are not enough to compensate their difference with the pre-trained model. We also found very important the mean subtraction step since the pre-trained model was trained with faces with some background and the canonically normalized appearance contains a black background. Hence, subtracting the global pixel mean was making all those zeros to be non-zero and thus lower the performance. The solution was to subtract the per-pixel mean. The best score for the AUC metric, 89.9 is achieved by considering the so popular pre-processing step, as used in [11].

The last 2 rows in Table II show the results obtained by our model when it is trained with centrally cropped Procrustes aligned faces. With this different setting, the performance of the model is enhanced, only matched by the canonically normalized setting when heavy data augmentation was used (face deformations). The main reason to this gain is due to the fact that VGG_Faces is pretrained with millions of raw images.

A possible drawback of keeping the image background could be that the CNN is helped by non-facial information (such as the arms) to improve its performance. In order to verify that the model is ignoring the background and that it is using only face information, we performed a class saliency visualization as described by Simonyan et al. [63]. In Fig. 5 it can be seen the average saliency map compared to the mean face, and by comparing both pictures it can be seen that the network bases its decision looking at the face region, without using background information. The average saliency map has been obtained by computing the saliency map of all the images and averaging them. According to [63] the saliency map of an image can be thought as the magnitude of the derivative of the output $S_c$ with respect to the input image $I$, because the magnitude of the derivative indicates which pixels need to be changed the least to affect the output the most, and therefore those pixels correspond to the region of the image that the network is using to give its output. The derivative is computed as following:

$$w = \frac{\partial S_c}{\partial I},$$  

and the saliency map $M \in \mathbb{R}^{m \times n}$ for an image $I \in \mathbb{R}^{m \times n}$ is computed as:

$$M_{ij} = \max_c |w_{i(j, c)}|,$$

where $h(i, j, c)$ is the index of the element in $w$ that corresponds to the $i$-th row, $j$-th column and $c$-th colour channel value of the image $I$. As the saliency map does not have a color dimension, the maximum magnitude of $w$ across all colour channels is selected to create the map.

The UNBC-McMaster Shoulder Pain Expression Archive Database is unbalanced, meaning that there are a lot more frames labeled as zero pain than frames labeled with some level of pain. There is a total of 48398 frames coded with a pain intensity, 40029 of them being labeled as zero-pain intensity. This means that the 83.6% of the examples of the dataset belong to the same class, whereas only the other 16.4% examples have some level of pain [11]. As stated by the authors in [49] the results of the accuracy metric is influenced by the skew in the testing data, whereas the AUC metric is not affected that much. Therefore, to avoid providing a score which is influenced by the skew in the data set, in Table II the first two columns correspond to the accuracy and AUC obtained when the score is skew normalized to mitigate the effect of imbalanced data. The last two columns correspond to the scores obtained testing the models with an unbalanced distribution. In the same way as the authors in [49], to calculate the skew normalized scores shown in Table II, we undersample the majority class at test time. This means that we randomly choose a set of no-pain samples (the majority class) that has as many images as the pain class (the minority class). Then, the normalized scores provided are calculated based on those samples. As stated by [49] the results of the accuracy metric are influenced by the skew in the testing data, whereas the AUC metric is not affected that much. That is why the...
Pain recognition from facial gestures is a specific task within the broader task of facial expression recognition. In order to evaluate the effectiveness and robustness of our proposed method, we apply it to the task of emotion recognition from facial pictures. Facial expressions can show different human emotions such as anger, disgust or happiness [64] so the task of emotion recognition from pictures of faces can be approached as a facial expression recognition task. Our method for pain recognition can be adapted to perform facial expression recognition very easily. For pain detection we perform a regression task, i.e. predicting the pain intensity of a face picture. To switch to emotion detection, we must now perform a classification task. To do so, we changed the number of output units in the output layer of the CNN from 1 output unit to N, where N is the number of emotions we want to recognize in one-hot encoding. The loss function was also be changed to the cross-entropy error between the correct output $y$ and the predicted output $\hat{y}$ as defined by the equation 12:

$$E(y, \hat{y}) = \sum_{n=1}^{N} y_n \log(\hat{y}_n)$$  

The output of the network $\hat{y}$ is the result of applying the softmax function to the outputs of the last layer, and the true label $y$, which is the one-hot representation of the emotion label assigned to a sample. To test our method on emotion
recognition we used the Extended Cohn-Kanade Emotion Dataset (CK+) [42].

1) CK+ Dataset: The emotion recognition CK+ dataset [42] has 593 sequences of 123 subjects which are FACS coded at the peak frame. In each sequence, the subject face evolves from a neutral face to a peak facial expression. Only 327 of the sequences are labeled with one of the following seven emotions: anger, contempt, disgust, fear, happy, sadness, surprise. In Fig. 6 there is an example of a peak frame for each of the seven emotions present in the dataset. Following the trend in other works [43, 65, 66], we split the sequences into 10 subject-exclusive folds in order to perform a leave-one-fold-out cross-validation to test our method on this dataset. To make sure that the classes are evenly distributed among folds, the subjects are randomly separated into 10 groups. In the same way as in other works [45, 65] we select the last three frames of each sequence to train the CNN. To train the LSTM we must provide fixed-length sequential inputs, and as the videos vary in duration, from 10 to 60 frames approximately, we have chosen the length of the sequences to be 10. For each video, we generate three different sequences of length 10, each sequence ending in one of the last three frames. If there aren’t enough frames in the video to build a sequence of length 10, the first frame is repeated at the beginning of the sequence. The results provided for the CK+ dataset are obtained by training on 9 of the 10 folds and leaving one out for testing, and repeating the process until each fold has been used for testing at least one. The accuracy provided is the average within the 10 folds.

2) Results on CK+: We provide two results for the CK+ dataset, the baseline accuracy obtained by the emotion classifier built on top of the CNN and the accuracy obtained by the LSTM model. In Table VI a comparison of our method scores against other state-of-the-art procedures reported in the literature can be seen. The results shown in the table are from seven emotion classes: anger, contempt, disgust, fear, happy, sadness, and surprise. The confusion matrix of the predictions on the test folds can be seen at Fig. 7. Other works [67] provide scores for the eight class problem where the neutral emotion is added. We can not construct sequences ending in a neutral frame because the neutral frame is always the first one, so we do not provide results for this task.

### Table VI: Results on the CK+ dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned crop (Ours)</td>
<td>94.5</td>
</tr>
<tr>
<td>Aligned crop + LSTM (Ours)</td>
<td>97.2</td>
</tr>
</tbody>
</table>

### IV. Conclusions

Pain recognition has been proved to be an important task for health-care. In this work we have faced the task of binary pain recognition on facial images from the deep learning perspective achieving state-of-the-art results when compared to leave-one-subject-out setups. This, however, has also exposed the problem of stating which is the correct comparison methodology since results from other works have been provided in terms of accuracy, AUC, subject exclusive and non-exclusive settings. We believe subject-exclusive is crucial and thus, provided all the results computed this way. Our approach of training a deep CNN for pain-level estimation already provided good results, and we have proved that using an RNN to exploit the temporal relation between frames improves the results even more. By training a CNN end-to-end to perform pain-level estimation our approach obtained an AUC of 89.6, increasing up to 93.3 when that same CNN is used to the extract features to train the RNN. Moreover, we prove the generality of our method by obtaining an accuracy of 97.2% on the CK+ facial emotion recognition dataset, a competitive score when compared to the state-of-the-art (97.3% in [65]).

### References


Figure 6: Examples of emotion frames. This figure shows one frame of each of the seven emotions. From left to right: anger, contempt, disgust, fear, happiness, sadness, surprise.

Figure 7: Emotion detection confusion matrix. Confusion matrix for the task of emotion detection in the CK+ dataset for the seven classes.


Pan Rodriguez received the MSc degree in Artificial Intelligence from KU Leuven in 2015. He is currently a Ph.D. student at the Image Social Evaluation (ISE) Lab of the Computer Vision Center and the Universitat Autònoma de Barcelona, Catalonia Spain. His research interests focus on machine learning, pattern recognition, and computer vision.

Guillem Cucurull received the BSc degree in Computer Science from Universitat Autònoma de Barcelona (UAB) in 2016. He is currently a MSc student in Computer Vision and a research assistant at the the Image Sequence Evaluation (ISE Lab) research group of the Computer Vision Center and UAB. His research interests include computer vision and machine learning techniques for human motion and behaviour analysis.

Jordi González received the PhD degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB) in 2004. He is an associate professor in computer science at the Computer Science Department, UAB. He is also a research fellow at the Computer Vision Center, where he has co-founded three spin-offs (Cloud Size Services, Visual Tagging, Care Respite) and the Image Sequence Evaluation (ISE Lab) research group. His research interests include machine learning techniques for the computational interpretation of social images, or Visual Hermeneutics.
Josep M. Gonfaus received the PhD degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB) in 2012. He participated in various Pascal challenges. He co-founded a spin-off based on their research by applying computer vision and deep learning techniques for analyzing social media data (Visual Tagging). His main research interests are deep learning techniques to mimic human cognitive capabilities.

Kamal Nasrollahi received the M.Sc. degree in computer engineering from Tehran Polytechnic, in 2007, and the Ph.D. degree in electrical engineering from Aalborg University, in 2010, with a focus on computer vision. He is currently an Associate Professor with the Visual Analysis of People Laboratory, Aalborg University, Denmark. He has been involved in five international research projects. His research interests include facial analysis systems, biometrics recognition, soft biometrics, and inverse problems. He has received an IEEE Conference Best Paper Award.

Thomas B. Moeslund (M’12) received the M.Sc.E.E. and Ph.D. degrees from Aalborg University, Aalborg, Denmark, in 2003 and 1996, respectively. He is head of a computer vision group at Aalborg University. His research is focused on all aspects of computer vision and image analysis. He has been involved in 12 national and international research projects, both as coordinator, WP leader, and researcher. He is reviewer for all major journals within the field. He serves as associate editor and editorial board member for four international journals. He has coedited one special journal issue, cochaired four international workshops/tutorials, and acted as PC member/reviewer for a number of conferences. He has published 90 peer-reviewed journal and conference papers.

F. Xavier Roca received the Ph.D. degree in computer science from Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain, in 1990. He is an Associate Professor and the Director of the Department of Computer Science, UAB. He is also a Research Fellow with the Computer Vision Center. He has been a Principal Researcher in several projects (public and private funds). He is working in technological transfer computer vision. The topics of his research are active vision, biometrics and tracking.