

Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video

Jing Zhou^{1,2}, Xiaopeng Hong^{*1}, Fei Su², and Guoying Zhao¹

¹Faculty of Information Technology and Electrical Engineering, the University of Oulu

²Beijing University of Posts and Telecommunications

1522670460@qq.com, xhong@ee.oulu.fi, sufai@bupt.edu.cn, gyzhao@ee.oulu.fi

Abstract

Automatic pain intensity estimation possesses a significant position in healthcare and medical field. Traditional static methods prefer to extract features from frames separately in a video, which would result in unstable changes and peaks among adjacent frames. To overcome this problem, we propose a real-time regression framework based on the recurrent convolutional neural network for automatic frame-level pain intensity estimation. Given vector sequences of AAM-warped facial images, we used a sliding-window strategy to obtain fixed-length input samples for the recurrent network. We then carefully design the architecture of the recurrent network to output continuous-valued pain intensity. The proposed end-to-end pain intensity regression framework can predict the pain intensity of each frame by considering a sufficiently large historical frames while limiting the scale of the parameters within the model. Our method achieves promising results regarding both accuracy and running speed on the published UNBC-McMaster Shoulder Pain Expression Archive Database.

1. Introduction

Measuring or monitoring pain intensity is crucial in pain medication, treatment or diagnosis to individuals who are unable to communicate verbally, such as newborns and patients in intensive care units. Normally, pain intensity measurements are conducted via self-report or checked by medical staffs (e.g., nurse or physician). But these measurements may cause unreliability or a large workload of hospitals. Thus, a reliable automatic pain intensity estimation model provides a more economical option to measure pain intensity of different subjects.

In the past decade, a plenty of approaches have been proposed for automatic pain intensity estimation. Table 1 provides a brief summary of typical approaches. Early researches tend to focus on estimating whether the subject is painful or not, and thus, conduct pain intensity estimation as a classification problem [2], [23], [21], [22], [24], [19], [27], [32].

More recently, an increasing number of researchers realize that simply judging whether it is painful or not for a whole sequence is too rough for fine-grained pain intensity estimation in practice. Therefore, they start to study frame-level pain intensity estimation and regard it as a regression problem.

One crucial issue here is to provide enough data where each frame is well labeled under a standard scientific measure to facilitate related researches. In 2008, Prkachin and Solomon [31] proposed a measure of pain intensity termed by Prkachin and Solomon Pain Intensity (PSPI) based on Facial Action Coding System (FACS) [8], [30]. PSPI is defined as a function of the intensity of six pain related Facial Action Units (AUs), which describe a set of facial configurations related to pain such as nose wrinkling and cheek-raising. By using PSPI as the frame-level intensity measure, a few recent works have been proposed for pain intensity regression. Kaltwang *et al.* [18] compared three approaches by using the locations of 66 facial landmark points, DCT, and LBP, as well as the combinations among them. Florea *et al.* used the histogram of topographical features and SVM, achieving a great result of average mean squared error (MSE) [11]. In [16], Hong *et al.* applied a second-order standardized moment average pooling (2Standmap) method which beats all approaches that only rely on a single descriptor.

However, traditional static features like LBP and DCT, which are extracted from separate frames, have inevitable limitations in describing relevant dynamic information re-

*Corresponding author

Feature descriptors	Pain levels	Measures of pain intensity	Classifier	Cross Validation
C-APP + S-PTS [2]	C-2	OPI, PSPI	SVM	Leave One Subject Out
PTS + APP [22]	C-2	PSPI	SVM	Leave One Subject Out
PTS, APP [21]	C-2	PSPI	SVM	Leave One Subject Out
SAPP +SPTS + CAPP [22]	C-2	PSPI	SVM+LLR	Leave One Subject Out
AAM [24]	C-2	OPI, PSPI	SVM	Leave One Subject Out
PLBP, PHOG [19]	C-2	PSPI	SVM	10-fold
Auto Encoder [27]	C-2	PSPI	SVM	Leave One Subject Out
TPS [32]	C-2	PSPI	DML + SVM	Leave One Subject Out
Canny Edge [15]	C-2/C-8	OPI, PSPI	TBM	3-fold
LBP [6]	C-2	PSPI	Transfer Learning	Leave One Subject Out
PCA [1]	C-3	VAS	SVM, Angular Distance	10-fold
DCT + LBP [18]	R	PSPI	RVR	Leave One Subject Out
Hess + Grad + AAM [11]	R	PSPI	SVM	Leave One Subject Out
2Standmap [16]	R	PSPI	RVR	Leave One Subject Out

Table 1: A brief summary of the recent methods proposed for pain detection and pain intensity estimation.

1st column - Feature descriptors: S-PTS: Similarity Normalized Shape, S-APP: Normalized Appearance, C-APP: Canonical Appearance, PTS: Normalized Shape, APP: Appearance, DCT: Discrete Cosine Transform, LBP: Local Binary Pattern, AAM: Active Appearance Model, PLBP: Pyramid LBP, PHOG: Pyramid Histogram of Orientation Gradients, TPS: Thin Plate Spline, PCA: Principal Component Analysis, Hess: Hessian based histograms, Grad: Gradient-based histograms; 2nd column - Pain levels: C: classification, R-n: n-level regression; 3rd column - Measures of pain intensity: OPI: Observer Pain Intensity, PSPI: Prkachin and Solomon Pain Intensity, VAS: Visual Analog Scale; 4th column - Classifier: SVM: Support Vector Machine, RVR: Relevance Vector Regression, NN: Nearest Neighbor, LLR: Linear Logistic Regression, TBM: Transferable Belief Model; and last column - Manner of Cross Validation.

quired by pain intensity estimation. For example, subjects tend to close eyes when they are suffering pain, but traditional features and static methods cannot differentiate between normal eye blink or eye closure that related to pain from independent frames. It thus results in unstable changes and peaks of the estimation among adjacent frames.

To overcome this problem, we attempt to encode the video not only from the separate frames but also among adjacent frames. In this paper, we propose a regression framework based on Recurrent Convolutional Neural Network (RCNN) for automatic frame-level pain intensity estimation. In the first step, we used Active Appearance Model (AAM) to track faces and warped all facial images of different poses. In the second step, given the vector sequences of the warped facial images, we used a sliding-window strategy to achieve fixed-length input samples of the recurrent network from the video sequence. Finally, we carefully design the architecture of the recurrent convolutional neural network for continuous-valued pain intensity. The proposed end-to-end pain intensity regression framework can predict the pain intensity of each frame by considering a sufficiently large historical frames while limiting the scale of the parameters within the model.

The main contribution of this work is that we propose an RCNN based framework to estimate pain intensity automat-

ically. According to the best knowledge of the authors, it is the first time that the recurrent (convolutional) neural network is applied to the task of pain intensity estimation. Correspondingly, the RCNN is used as an end-to-end regressor, which outputs continuous scores rather than discrete labels as in the problem of classification.

The proposed regression network is evaluated on the published UNBC-McMaster Shoulder Pain Expression Archive Database, where our method gets promising results with a real-time testing speed.

The remaining content of this paper is organized as follows. Section 2 briefly introduces the background of RCNN. Section 3 details the proposed framework. Quantitative experimental results are provided in Section 4. Section 5 concludes the paper.

2. Recurrent Convolutional Neural Network

In the past few years, Convolutional neural network (CNN) has made a great success in various computer vision tasks, such as image classification [33], object detection [12], and tracking [35]. CNN has been characterized by local connections, weight sharing, and local pooling, which largely attribute to excellent performances. Recurrent neural network (RNN) has a long history in the artificial neural network community [5], [9], [10], the most successful

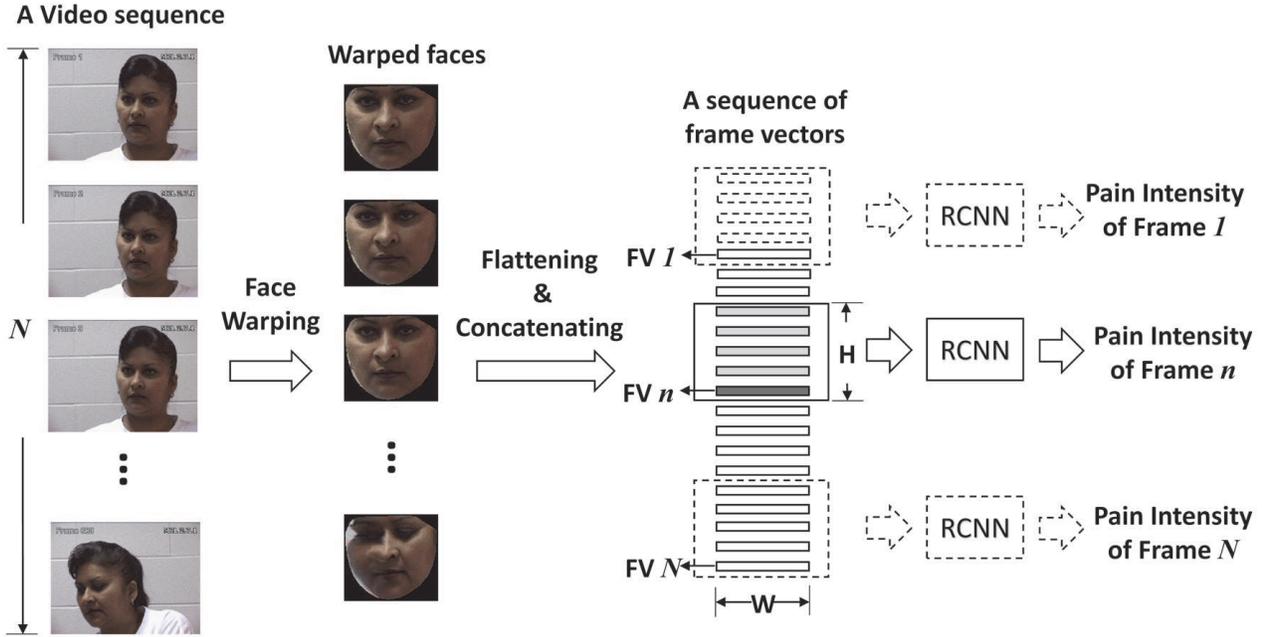


Figure 1: The framework of the proposed pain intensity estimation approach.

applications refer to sequential tasks such as [13], [14]. RNN has been characterized by connecting hidden layers of the current time step and several previous time steps. Because RNN reserves the temporal information in sequences, it achieves a great performance in sequential tasks. Combining the advantages of CNN and RNN, different structures of networks were proposed to fuse convolutional layers and recurrent layers to capture relevant contextual information from raw pixels in static images. In 2014, Pinheiro and Collobert [29] used extra recurrent connections from the top layer to the bottom layer of a CNN for scene labeling. In [20], Liang and Hu proposed an RCNN for object recognition by using recurrent connections within the same layer. Their models are different from our proposed RCNN regression, which lies in two folds: first, the RCNNs in [29] and [20] are applied to the tasks based on static images while here we used RCNN for modeling the temporal information in videos; secondly, their RCNNs are used as classifiers by using softmax function as the activation function of the fully connected layer, ours is used as a regressor for estimate pain intensity. The architecture of normal RCNN will be explained in Section 3.

3. Frame-by-Frame Regression Network

3.1. The Framework

The key problems of pain intensity estimation can be summarized as four blocks. Firstly, each incoming facial

frame (n) of the testing video sequence should be aligned and warped to the same frontal pose. Secondly, in order to keep spatial and temporal information at the same time, we need to convert each warped face into a (3-channel (RGB)) frame vector (FV_n). Thirdly, because of the fixed height (H) of the input that our RCNN requires, we applied a sliding window to achieve testing samples. When testing the frame n , the testing sample contains several continuous adjacent frame vectors before the frame n (padding zeros if $n < H$). Finally, we fed the samples to a trained RCNN, and the network will output the PSPI predictions frame by frame. The whole framework is shown as Fig. 1.

As for the training process, we used a random strategy to achieve the training samples of fixed length. Similarly, we converted all frame images into frame vectors. All converted frame vector sequences will be immersed in a training pool; then the network uses windows to select randomly a subset of the training data to conduct one training iteration. The length of every training sample (H) indicates the number of continuous frames that the recurrent network will use at one time. Then, these training samples will be fed into the RCNN regression structure to start learning.

3.2. Preprocessing

A pain intensity estimation algorithm should be both robust to face pose and the identity of the subject (not subject dependent). To achieve invariance to different face poses, we exploited an Active Appearance Model (AAM)

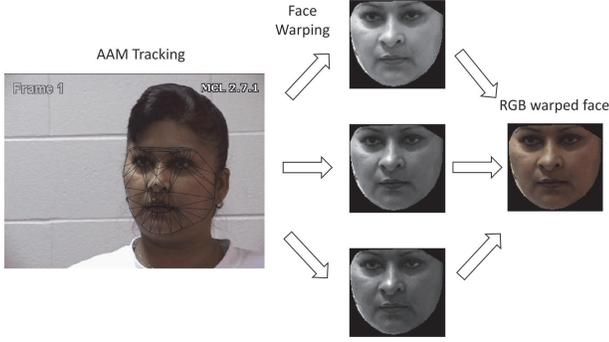


Figure 2: AAM tracking; R, G, B warped faces; and RGB warped face.

to warp all facial images of different poses into the same frontal pose. AAM tracks the face and extracts visual features, finding the key points on faces, such as eyebrows, and the outline of faces. These AAM landmark points constitute many non-overlapping triangles, which can warp and align different faces into the same 2D triangulated mesh after some linear shape variation [7], [34]. In the process of face aligning and warping, we used the same facial triangulated mesh for all subjects. We warped every facial image in RGB channels separately, then combined all channels back to get the final RGB warped faces (see Fig. 2).

The input samples of our RCNN structure should be no more than two dimensions, but to reserve the temporal information among frames and the spatial pixel information of warped facial images at the same time, we considered some different ways to convert each frame into a 1D vector, such as flattening or extracting feature vectors. Finally, it turned out that flattening is an effective way though it may lose some structural information of the images. After flattening, we concatenated all 1D flattened warped facial images in frame order to achieve frame vector sequences.

3.3. Architecture of RCNN

The basic idea of RCNN is to add recurrent connections within every convolutional layer of the feed-forward CNN [20]. The overall architecture of RCNN is shown in Fig. 3. The first layer ($C1$) is the standard feed-forward convolutional layer without recurrent connections. Following ($C1$), there are several recurrent convolutional layers ($RCL1 \sim RCLm$), with a max pooling layer between every two RCLs. Normally, the final output layer is a softmax layer in the tasks of classification.

Each RCL is constituted by several iterative convolutions, sharing weights in hidden layers among $T + 1$ time steps. If unfolding an RCL, the layer can be seen as a feed-forward subnetwork with the depth of $T + 1$ (see Fig. 4). The difference between an RCL and a $(T + 1)$ -layer CNN

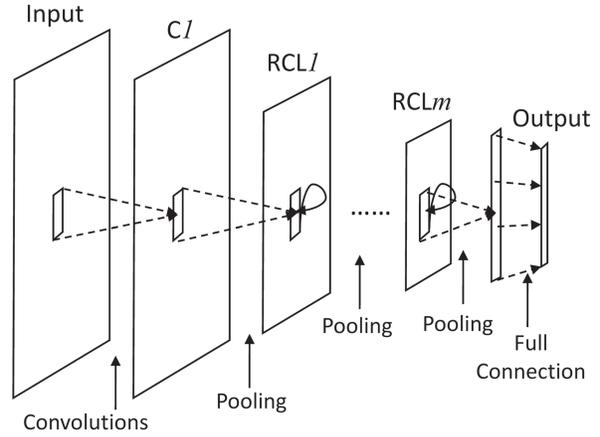


Figure 3: The overall architecture of RCNN.

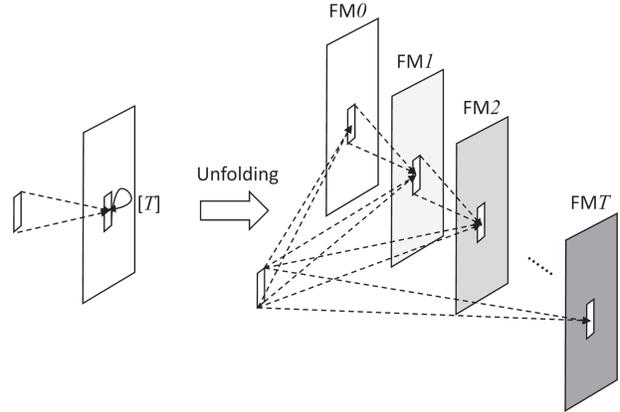


Figure 4: Unfolding an RCL.

is that the inputs of RCL are all the values of time steps from 0 to T , but the inputs of CNN are the values of one fixed time step. Thus, unfolding an RCNN through time steps in RCLs can result in an arbitrarily deep network with a fixed number of parameters.

The overall depth of the model is crucial for obtaining good performance [37]. The existence of deeper layers or longer paths among layers in a network makes it possible for the network to learn highly complex features. On the contrary, shorter paths may help gradient backpropagation during training. RCNN is actually a CNN with flexible paths between the input layer to the output layer, which expands the depth of the network but also facilitates the learning [20]. In the structure of an RCL, there are several paths from the first feature map of convolution ($FM0$) to the last feature map (FMT). In Fig.4, the darker the feature map is, the deeper the path is. Attribute to the iteration in an RCL,

the length of path ranges from 1 to $T + 1$, including the first path of the convolutional layer. In our framework, we used four ($m = 4$) RCLs in the whole RCNN architecture. Therefore, the length of the iterative path will range from 6 to $4(T + 1) + 2$ including the first path of CI and the last path of the output layer. The length of recurrent time steps (T) was empirically set as 3.

The following subsection will introduce how the output layer is modified for continuous-valued predictions. And more detailed implementation setup of our network will be described in Section 4.

3.4. Continuous Predictions

The recurrent convolutional neural network (RCNN) are usually used to solve classification problems such as image classification [20] and scene labeling [29]. Correspondingly, to assign feature vectors to one of the C categories, the final output layer is a *softmax* layer whose output is given by:

$$\hat{y}_i = \frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{i'} \mathbf{w}_{i'}^T \mathbf{x}}, \quad (1)$$

where \hat{y}_i is the predicted probability belonging to the i th category, for $i = 1, 2, \dots, C$, and \mathbf{x} is the feature vector generated by the global max pooling before the output layer. The training process is performed by minimizing the cross-entropy loss function as:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]. \quad (2)$$

As for the estimation of pain intensity, the network should allow continuous-valued predictions. A linear function is therefore simply used as the activation function in the output layer of the network:

$$\hat{y} = \mathbf{w}^T \mathbf{x}, \quad (3)$$

where \hat{y} is the continuous predicted value of the network, and \mathbf{x} is the feature vector. Correspondingly, the loss function is modified to the mean squared error function as:

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2, \quad (4)$$

rather than the cross-entropy function in Eq. 2. With it, the output becomes continuous so that it turns a regressor. Training is performed by minimizing the MSE function using the back-propagation through time (BPTT) algorithm [36]. This is equivalent to using the standard BP algorithm on the time-unfolded network. The final gradient of a shared weight is the sum of its gradients over all time steps.

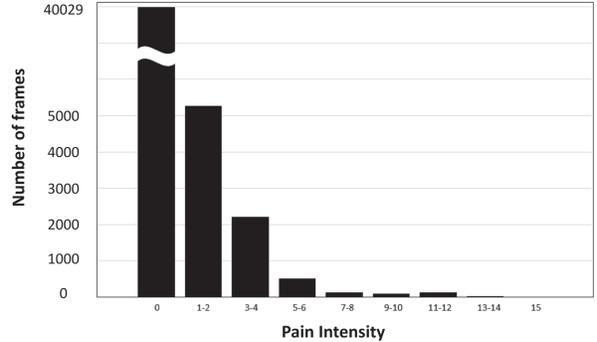


Figure 5: Frame distribution of the PSPI (0-15).

4. Experiments

4.1. Pain Intensity Dataset

Recently, researchers at the McMaster University and University of Northern British Columbia (UNBC) published a shoulder pain expression archive database [25]. This database is the most common database to be used to assess pain detection or pain intensity estimation methods. The database captured face videos of subjects (66 females and 63 males) when they were performing a series of active and passive range-of-motion tests to their affected and unaffected limbs on two separate occasions. Out of which videos of active tests are publicly available for research purposes. In this database, each video was coded by FACS in frame level. Observer and self-report measurements in sequence level were also taken. The PSPI score was computed to quantify pain intensity in 16 discrete levels (0-15) based on AUs [8], [30]. In this paper, we used the videos of active tests to perform pain intensity estimation experiments, with the 16-level PSPI as the ground-truth. Active tests include 200 sequences of 25 subjects, with totally 48,398 frames of 320×240 pixels. We noticed that the frame distribution of the PSPI is quite unbalanced as shown in Fig. 5.

To solve the unbalanced training samples of 16 levels, we designed a weighted strategy to keep training samples of all labels balanced to some extent. The network selects a subset of the training samples randomly to conduct one training iteration. The subset contains samples of all PSPI levels, and the percentage of samples corresponding to each PSPI level is weighted manually.

4.2. Measurement

In our experiments, we conducted a leave-one-subject-out strategy which leads to 25-fold cross-validation to assess our method. We left all sequences of one chosen subject as the testing set and the rest sequences of 24 subjects as the training set at the same time. The average Mean Squared

Error (MSE) and Pearson Product-moment Correlation Coefficient (PCC) were calculated by:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2, \quad (5)$$

$$PCC = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (6)$$

where N is the total number of frames of testing sequences. y_i and \hat{y}_i are the ground-truth and the pain intensity estimation of the i frame, respectively. $\bar{\hat{y}}$ and \bar{y} are the sample mean of $\{y_1, \dots, y_N\}$ and $\{\hat{y}_1, \dots, \hat{y}_N\}$.

4.3. Implementation Details

As is described in Section 3, we got 3-channel (RGB) frame vector sequences ($H \times W$) as the input of the network. The choice of H is strongly related to the time cycle of the pain occurrence. In our experiments, H and W were empirically set as 30 and 713, respectively. In each RCL, we used one convolutional layer first (functioning as a feed-forward layer), then connected three iterations ($T = 3$ in Fig.4) following the feed-forward layer. In the fully connected layer, we used a linear function as the activation to conduct the regression task and the MSE function as the loss measurement. A summary of the main network configurations is shown in Table 2.

Layer type	Configurations
Input	$W(713) \times H(30) \times 3$ RGB vector sequence
Convolution 1	maps:256, $k : 3 \times 3, s : 1$
Max pooling 1	$p : 4 \times 1, s : 4 \times 1$
RCL 2	feed-forward map:256 $k : 1 \times 1, s : 1$ 3 iteration maps:256 $k : 3 \times 3, s : 1$
Max pooling 2	$p : 4 \times 1, s : 4 \times 1$
RCL 3	feed-forward map:256 $k : 1 \times 1, s : 1$ 3 iteration maps:256 $k : 3 \times 3, s : 1$
Max pooling 3	$p : 4 \times 4, s : 4 \times 4$
RCL 4	feed-forward map:256 $k : 1 \times 1, s : 1$ 3 iteration maps:256 $k : 3 \times 3, s : 1$
Max pooling 4	$p : 2 \times 2, s : 2 \times 2$
RCL 5	feed-forward map:256 $k : 1 \times 1, s : 1$ 3 iteration maps:256 $k : 3 \times 3, s : 1$
Max pooling 5	$p : 1 \times 1, s : 1 \times 1$
Output	$H(30)$ predictions

Table 2: A summary of the main network configurations. k, s and p stand for the kernel size, stride and pooling size in the related layer, respectively.

The initial learning rate was set heuristically and annealed according to a schedule pre-determined on the cross-validation set. When the accuracy improved so slowly, we decreased the learning rate to its 1/10. Annealing was used three times through a whole training process so that the final learning rate was 1/1000 of the initial value. The momentum was fixed at 0.9. Weight decay decreased overfitting as well as dropout. Moreover, we used a batch normalization technique [17] following the first convolutional layer and every feed-forward layer in RCLs to accelerate the training process. We implemented the network within the Theano 0.8 [3], [4] framework. Our experiments were carried out on a workstation with two 2.30GHz Intel(R) Xeon(R) E5-2650 v3 CPU, 320GB RAM, and an NVIDIA(R) Tesla K80 GPU to run our experiments. The average testing time is 25 frame per second.

4.4. Experimental Results

In our experiments, we compared our method with the state-of-the-arts on the UNBC-McMaster Shoulder Pain Expression Archive Database as shown in Table 3.

Methods	MSE	PCC
PTS [18]	2.59	0.36
DC [18]	1.71	0.55
LBP [18]	1.81	0.48
(DCT+LBP)/RVR [18]	1.39	0.59
2Standmap [16]	1.42	0.55
Hessian Histograms [11]	3.76	0.25
Gradient Histograms [11]	4.76	0.34
Hess+Grad [11]	3.35	0.41
VGG-face CNN SVR	1.70	0.43
RCNN regression	1.54	0.65

Table 3: Comparison of the proposed approach with other approaches in the literature.

Single features, mean feature fusion, and RVR feature fusion were proposed in [18], which includes the combinations of DCT and LBP. The mean feature fusion method calculates the weighted mean of the responses of the regression function based on one single descriptor directly, and the RVR feature fusion method using Relevance Vector Regression. [11] extracts Hessian based histograms, gradient based histograms and AAM landmarks as features and uses SVM as the classifier, getting the best average MSE among all methods. [16] applies a second-order standardized moment average pooling (2Standmap) method which beats all approaches that only rely on a single descriptor. Additionally, we also used a method by extracting CNN features (VGG-face CNN SVR) as a baseline method of neural networks. We fed all warped facial images into the VGG-face

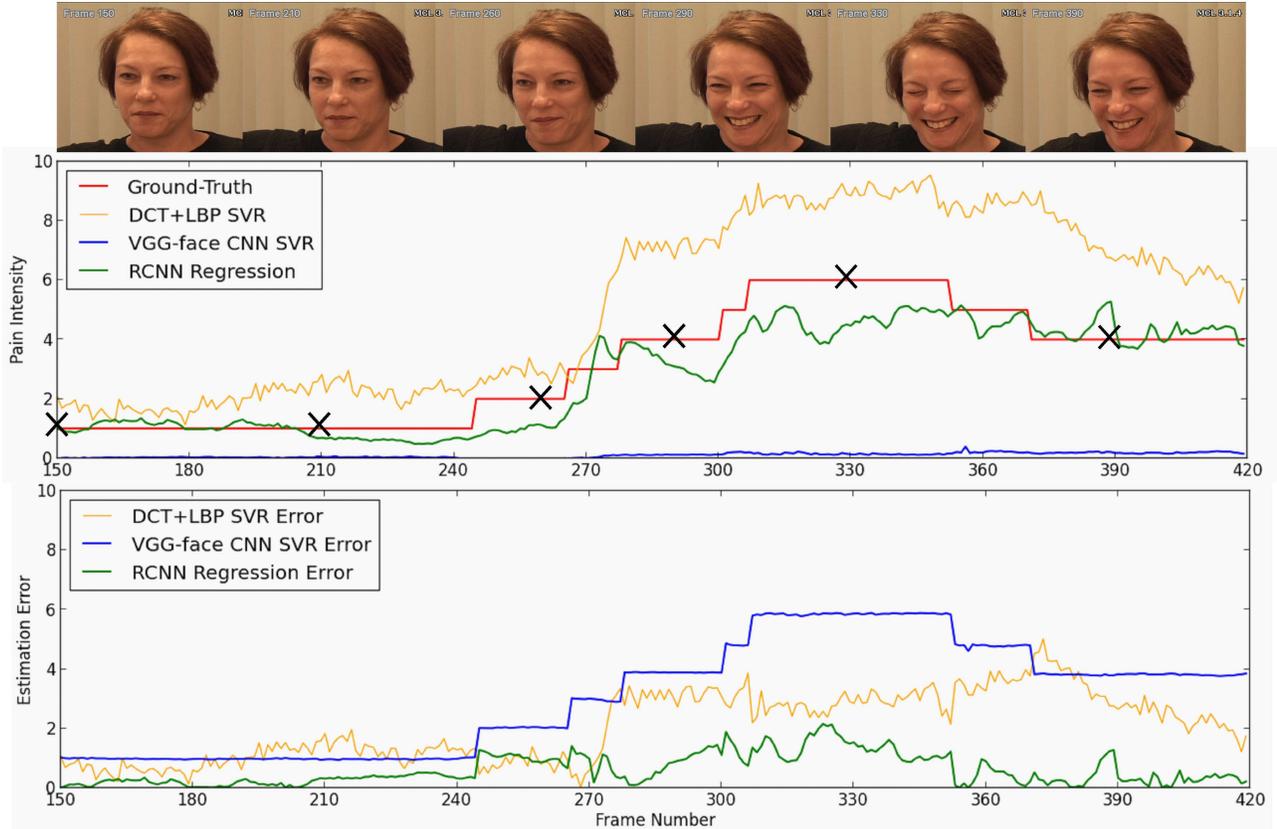


Figure 6: An example sequence of pain intensity and estimation error for DCT+LBP SVR, VGG-face CNN SVR, and RCNN Regression. (N.B. The "X" on the Ground-Truth correspond to the frames of number 150, 210, 260, 290, 330, and 390).

CNN [26], then we delivered the VGG-face descriptors to linear SVR [28].

As for our proposed method, we used regression RCNN to conduct pain intensity estimation. We got promising results of the average MSE and PCC of 1.54 and 0.65, respectively. It indicates that our method is effective. Regarding the computational speed, our method was able to process 25 frames per second on our workstation (two 2.30GHz Intel(R) Xeon(R) E5-2650 v3 CPU, 320GB RAM, and an NVIDIA(R) Tesla K80 GPU). Therefore, our method is testing efficient for real time application. Fig. 6 shows an example pain intensity estimation sequence (frame 150 to 420) of one subject using: DCT+LBP SVR, VGG-face CNN SVR, and our proposed RCNN regression. These methods got the MSE of 3.83, 10.06, 1.12 and the PCC of 0.90, 0.55, 0.89 respectively in this sequence. Compared to the other two methods, RCNN regression has a smoother approximation and smaller estimation error. Besides, from the frame 270 on, the subject appears to close her eyes. Normally, eye closure relates to pain to some extent. Using traditional structural features (*e.g.*, LBP and DCT) and static methods (trained per frame) cannot differentiate eye

blink (short time) and eye closure (long time), so the model tends to result in that all eye-closed images are strongly related to pain as it has learned in the training stage. It is the exact reason that the estimation of pain intensity by using DCT+LBP SVR keeps a continuous high level after the frame 270. However, our proposed regression RCNN is a dynamic method that predicts one frame by using several adjacent frames, which keeps the estimation line stable, smooth, and closed to the ground-truth.

5. Conclusion

In this paper, we propose an automatic frame-by-frame pain intensity estimation framework in video based on a regression recurrent convolutional neural network. By leveraging the RCNN, firstly, the proposed framework predicts the pain intensity of each frame by considering a sufficiently large historical frames while limiting the scale of the parameters within the model; secondly, the framework encodes the spatial information, without losing temporal information of videos. To achieve continuous pain intensity estimation frame by frame, we modify the loss and the acti-

vation functions in the last fully connected layer of normal RCNN so that it has an output of continuous values. The proposed method is evaluated the UNBC-McMaster Shoulder Pain Expression Archive Database. The comparisons with state-of-the-art methods are promising. We also show that the output of the proposed method turned out stable, smooth, and also can avoid unstable jumps or peaks among frames which are inevitable via static methods. Last but not least, our method is computationally efficient for real-time applications. Future work may study accelerating the training section of the RCNN for pain intensity estimation.

Acknowledgment

This work is sponsored by the Academy of Finland, Infotech Oulu and Tekes Fidipro Program. Moreover, Xiopeng Hong is partly supported by the Natural Science Foundation of China under the contract No. 61572205. Finally, we appreciate Mr. Ming Liang for sharing the codes of the recurrent convolutional neural network.

References

- [1] M. Adibuzzaman, C. Ostberg, S. Ahamed, R. Povinelli, B. Sindhu, R. Love, F. Kawsar, and G. M. T. Ahsan. Assessment of pain using facial pictures taken with a smartphone. In *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual*, volume 2, pages 726–731. IEEE, 2015.
- [2] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face—pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009.
- [3] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements; 2012. In *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*.
- [4] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, page 3. Austin, TX, 2010.
- [5] G. A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.
- [6] J. Chen, X. Liu, P. Tu, and A. Aragonés. Person-specific expression recognition with transfer learning. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2621–2624. IEEE, 2012.
- [7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.
- [8] P. Ekman, W. Friesen, and J. Hager. Facial action coding system: Research nexus. 2002. *Network Research Information, Salt Lake City, UT, USA*.
- [9] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [10] B. Fernandez, A. G. Parlos, and W. K. Tsai. Nonlinear dynamic system identification using artificial neural networks (anns). In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pages 133–141. IEEE, 1990.
- [11] C. Florea, L. Florea, and C. Vertan. Learning pain from emotion: transferred hot data representation for pain intensity estimation. In *Computer Vision-ECCV 2014 Workshops*, pages 778–790. Springer, 2014.
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [13] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):855–868, 2009.
- [14] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.
- [15] Z. Hammal and M. Kunz. Pain monitoring: A dynamic and context-sensitive system. *Pattern Recognition*, 45(4):1265–1280, 2012.
- [16] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikäinen. Capturing correlations of local features for image representation. *Neurocomputing*, 184:99–106, 2016.
- [17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [18] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012.
- [19] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz. Pain detection through shape and appearance features. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.
- [20] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3367–3375, 2015.
- [21] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M. Prkachin. Automatically detecting pain using facial actions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. AII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009.
- [22] P. Lucey, J. F. Cohn, S. Lucey, S. Sridharan, and K. M. Prkachin. Automatically detecting action units from faces of pain: Comparing shape and appearance features. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 12–18. IEEE, 2009.
- [23] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting

- pain in video through facial action units. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(3):664–674, 2011.
- [24] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.
- [25] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011 *IEEE International Conference on*, pages 57–64. IEEE, 2011.
- [26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 1(3):6, 2015.
- [27] H. Pedersen. Learning appearance features for pain detection using the unbc-mcmaster shoulder pain expression archive database. In *Computer Vision Systems*, pages 128–136. Springer, 2015.
- [28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [29] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *International Conference on Machine Learning*, pages 82–90, 2014.
- [30] K. M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, 1992.
- [31] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008.
- [32] N. Rathee and D. Ganotra. A novel approach for pain intensity detection based on facial feature deformations. *Journal of Visual Communication and Image Representation*, 33:247–254, 2015.
- [33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [34] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 593–600, 2013.
- [35] N. Wang, S. Li, A. Gupta, and D.-Y. Yeung. Transferring rich feature hierarchies for robust visual tracking. *arXiv preprint arXiv:1501.04587*, 2015.
- [36] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.
- [37] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV 2014*, 2014.