ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Hierarchical Combination of Video Features for Personalised Pain Level Recognition

Patrick Thiam, Viktor Kessler and Friedhelm Schwenker *

Ulm University - Institute of Neural Information Processing
James-Franck-Ring 89061 Ulm - Germany

Abstract. In this work, we present a personalised participant independent pain recognition system based on the video channel. Instead of using an entire annotated dataset to train a classification model that would be later applied to an unseen participant, a similarity metric is used to select the most interesting annotated samples based on the data of the unseen participant. These samples are subsequently used to train a model adapted to the unseen participant. The selection process helps to avoid redundant and irrelevant data samples, thus improves the performance as well as the efficiency of the trained model. From the video channel, several features are extracted and subsequently fed into an hierarchical fusion architecture to further improve the performance of the system.

1 Introduction

Countless factors as old age, mental impairment or neurodegenerative disorders could affect an individual's capability of pain perception and assessment. A reliable and automatic pain assessment system would provide valuable insights in such cases, and would help to define an appropriate therapy that could considerably improve the quality of life of the patients. A huge variety of approaches for automatic pain recognition has been developed in the last decades. Since the facial region is used to convey a huge amount of information about an individual's affective state, almost all approaches involved the analysis of facial expressions [1, 2, 3]. In the present work we present a personalised pain recognition system in a participant independent scenario based on the analysis of facial expressions. The goal behind the personalisation setting is to try to improve the performance of a classification model on a dataset specific to an unseen participant, by choosing appropriate samples from the labelled set to train the model. This selection process should help to avoid redundant and irrelevant data samples and to improve the efficiency as well as the performance of the trained model. The whole personalisation setting is embedded within an hierarchical fusion architecture, developed to take advantage of the multitude of features that can be extracted from the facial region.

*We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This paper is based on work done within the project *SenseEmotion* funded by the Federal Ministry of Education and Research (BMBF). The work of Viktor Kessler and Friedhelm Schwenker is supported by the SFBTRR 62 funded by the German Research Foundation (DFG).

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

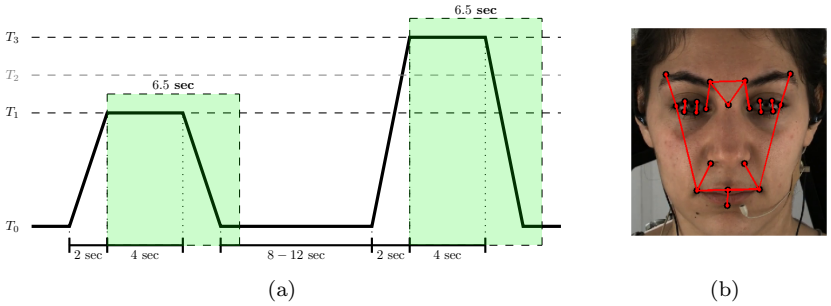

Fig. 1: Experimental settings and video features extraction. **Left:** Pain stimulation. T_0: baseline temperature (32°C); T_1: pain threshold temperature; T_2: intermediate temperature; T_3: pain tolerance temperature. The features are extracted from the green windows of length 6.5 seconds. **Right:** Facial region and distances computed between the tracked facial landmarks.

2 Dataset and Feature Extraction

The data used for the evaluation of the developed system consists of 40 participants (20 male and 20 female), each subjected to two sessions of pain stimulation. Each session had a length of about 40 minutes and involved several pain elicitations using a Medoc Pathway thermal simulator[1]. During the experiments, the participant's demeanour was recorded through several channels including audio, high resolution video and bio-physiology (electromyographic activity of the trapezius muscle (EMG), galvanic skin response (GSR), electrocardiogram (ECG), respiration (RSP)). For each session, the sensors recording the galvanic skin conductance were always attached to the left hand, while the thermode used for the heat elicitation was attached from a session to the next to a different forearm (right and left).

Before the data was recorded, each participant's pain threshold temperature (T_1) and pain tolerance temperature (T_3) were determined. Based on these temperatures, an intermediate heat stimulation temperature (T_2) was computed such that the range between both the threshold and tolerance temperatures was divided into 2 equally spaced ranges. The baseline temperature (T_0) corresponding to no heat elicitation was set at a constant temperature of 32°C. Each temperature was elicited randomly 30 times with a pause of 8 to 12 seconds between consecutive stimuli (the higher the temperature, the longer the pause). Each stimulation consisted of a 2 seconds onset during which the temperature was gradually elevated starting from T_0 until the targeted temperature was reached. Subsequently, the latter was maintained for 4 seconds before being gradually dropped until the baseline temperature was reached (see Fig. 1(a) for more details).

The facial behaviour analysis toolkit OpenFace [5] is used to automatically de-

[1] http://medoc-web.com/products/pathway-model-ats/

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

tect the facial region as well as a set of $2D$ facial landmarks (see Fig. 1(b)) in each frame of the video recordings. Furthermore, using the same tool, estimates of the head pose consisting of 3 rotation angles and 3 head position parameters were also extracted. Based on these parameters, a set of distances (see Fig. 1(b)) was computed from each frame in order to capture the deformation of the facial area at the frame level. Each of the computed distances as well as the head rotation angles and the head position parameters yields a signal for an entire window (see green boxes in Fig. 1(a)). These signals are low-pass filtered and the first and second derivatives of the filtered signals are computed. A set of 14 functionals (mean, median, maximum, minimum, range, standard deviation, kurtosis, skewness, first and second quartile, inter quartile, 1%-percentile, 99%-percentile, range of 1%-percentile and 99%-percentile) are subsequently applied on the latter to extract several statistical parameters that are used as geometric and head pose related features.
Furthermore, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) and Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) features were extracted. Prior to the extraction, each detected facial region within the defined window was divided into a 4×4 grid of cells with a 25% overlap from one cell to the next. The window itself was divided in 3 overlapping segments of 2.5 seconds each with an overlap of 0.5 seconds from one segment to the next. Both LGBP-TOP and LBP-TOP operators were applied on each resulting cuboid. In both cases, the generated histograms were subsequently concatenated to form the window level feature vector. Finally, Pyramid Histogram of Oriented Gradients (PHOG) features were also extracted from the facial region in each frame. The feature for the whole window was generated by performing a max pooling from the frame level feature vectors.

3 Hierarchical Fusion Architecture

In order to take advantage of the diversity of the features extracted from the facial region, a hierarchical fusion architecture depicted in Fig. 2 is designed. Each feature goes through a dimensionality reduction step consisting in discarding features with low variance, as well as highly correlated features before being fed into the classification system. The performance of the designed system is assessed in a leave one participant out setting and by focussing on the classification task T_0 vs T_3. The training set is divided into three subsets. The first subset is used to train a first layer of Random Forests classifiers, each trained on one specific feature set. The second subset is used to train a pseudo-inverse mapping as well as a multi-layer perceptron (MLP) mapping, using the scores generated by each Random Forests classifier of the first layer. The MLP mapping consists of a single hidden layer with 150 neurons, each with a log-sigmoid transfer function. Using both generated layers and the third subset, a final pseudo-inverse mapping is trained on the scores of both pseudo-inverse and MLP fusion mappings.

The classification results for each dataset (left and right forearm) depicted in Table 1 and Table 2 prove the effectiveness of the designed fusion architecture.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

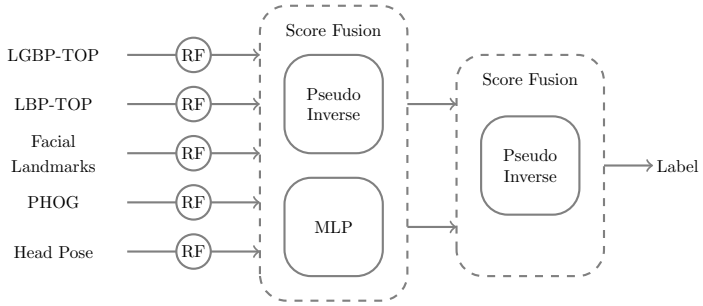

Fig. 2: Hierarchical fusion architecture.

Table 1: Leave one participant out cross validation performance for the classification task T_0 vs T_3. The averaged classification accuracy and the standard deviation over the entire 40 participants are depicted.

Dataset	LGBP-TOP	LBP-TOP	Landmarks	PHOG	Head Pose
Left Forearm	62.36% (±0.137)	60.19% (±0.122)	**65.42% (±0.149)**	59.09% (±0.114)	60.70% (±0.157)
Right Forearm	63.21% (±0.138)	63.71% (±0.137)	**64.94% (±0.166)**	61.48% (±0.135)	64.27% (±0.145)

The facial landmarks features outperform the other extracted features. The fusion performed by the pseudo-inverse mapping in the last layer is able to slightly improve the performance of the whole system in comparison to the results of the pseudo-inverse and the MLP mappings in the second layer. Still, the performance of the system is quite low (65.65% for the left forearm and 67.62% for the right forearm). This can be explained by the fact that some participants are unresponsive to the elicited pain stimuli and display no observable facial expression.

Table 2: Leave one participant out cross validation performance for the classification task T_0 vs T_3. The averaged classification accuracy as well as the standard deviation over the entire 40 participants are depicted.

Dataset	PINV	MLP	Final
Left Forearm	65.35% (±0.170)	63.45% (±0.146)	**65.65% (±0.162)**
Right Forearm	67.60% (±0.153)	66.68% (±0.151)	**67.62% (±0.150)**

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

4 Personalisation: Adaptive Participants Subset Selection

Based on the findings described in Section 3, the goal of the personalisation setting is to improve the performance of the system by carefully selecting the samples that are used to train the classifiers, based on the dataset specific to an unseen participant. The assumption is that the performance of the system can be improved by using samples from participants that are similar to an unseen participant. Therefore, similarity metrics have to be defined in order to proceed with the selection of such participants and furthermore improve the efficiency and the speed of generation of an optimal classification model. For the current work, the Hausdorff distance [6] is used as similarity measure. Given two finite sets $X, Y \subset \mathbb{R}^n$, the Hausdorff distance is defined as

$$\begin{aligned} &H(X,Y) = max(h(X,Y), h(Y,X)) \\ &where\ h(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\| \end{aligned} \tag{1}$$

and $\| \cdot \|$ is a norm on the points of X and Y (e.g. Euclidean norm). Two variants of the Hausdorff distance, one based on the average and the other on the median of the distances between the points of two sets, are also used:

$$avgH = \frac{1}{2}\left(\frac{1}{|X|}\sum_{x \in X} \min_{y \in Y} \|x - y\| + \frac{1}{|Y|}\sum_{y \in Y} \min_{x \in X} \|x - y\|\right) \tag{2}$$

$$medH = \frac{1}{2}(\underset{x \in X}{med}\{\underset{y \in Y}{min}\|x - y\|\} + \underset{y \in Y}{med}\{\underset{x \in X}{min}\|x - y\|\}) \tag{3}$$

These measures are computed to determine the similarity between the dataset of an unseen participant and the dataset of each of the participants in the training set. The results are subsequently sorted in increasing order of the similarity measure. The facial landmarks feature set are used for the selection of the participants. The hierarchical fusion architecture described earlier is used to perform the classification. Fig. 3 depicts the results of the experiments. No improvement can be observed by using the median based variant of the Hausdorff distance. Significant improvements of about 1.15% and 1.14% from the baseline performance can be achieved on the left forearm's dataset by selecting 30 participants and by using respectively the average based variant and the Hausdorff distance. With the same amount of participants, a slight improvement of about 0.18% from the baseline performance can also be achieved on the right forearm's dataset with the Hausdorff distance.

5 Conclusion

In this work, a personalised participant independent pain recognition system based on the video channel was presented. The personalisation method based on the Hausdorff distance is able to improve the performance of the system,

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

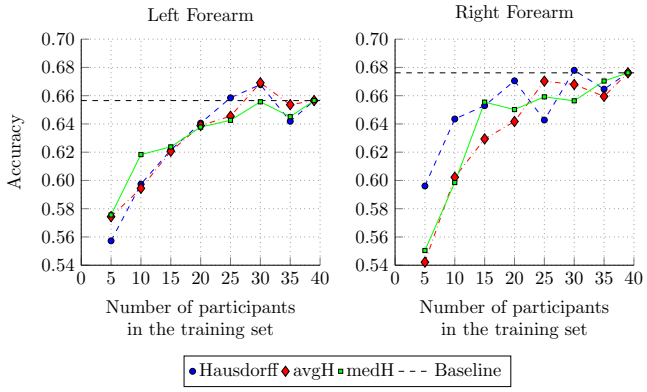

Fig. 3: Personalisation results. The baseline (dashed black line) corresponds to the average accuracy of the leave one participant out cross validation. The similarity measure based on the Hausdorff distance outperforms the baseline for both datasets with an average accuracy of respectively 66.79% and 67.80% by selecting 30 participants.

in comparison to a setting in which all of the available annotated datasets are used to perform the classification of an unseen participant. There is some room for improvement and several similarity measures or selection methods [4] should be experimented with in order to find an optimal subset of participants for the training of a model adapted to an unseen participant. Furthermore, an architecture should be designed that takes a multi-modal dataset into account.

References

[1] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368 – 377. Springer Berlin Heidelberg, 2012.

[2] Philip Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain recognition from video and biomedical signals. In *In Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 4582–4587, 2014.

[3] Markus Kächele, Mohammadreza Amirian, Patrick Thiam, Philipp Werner, Steffen Walter, Günther Palm, and Friedhelm Schwenker. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, pages 1–13, 2016.

[4] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, pages 854–864, 2016.

[5] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016.

[6] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 850–863, 1993.