Automated discomfort detection for premature infants in NICU using time-frequency feature-images and CNNs

Sun, Yue, Kommers, Deedee, Tan, Tao, Wang, Wenjin, Long, Xi, et al.

Yue Sun, Deedee Kommers, Tao Tan, Wenjin Wang, Xi Long, Caifeng Shan, Carola van Pul, Ronald M. Aarts, Peter Andriessen, Peter H. N. de With, "Automated discomfort detection for premature infants in NICU using time-frequency feature-images and CNNs," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 113144B (16 March 2020); doi: 10.1117/12.2549250

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Automatic Discomfort Detection for Premature Infants in NICU Using Time-Frequency Feature-Images and CNNs

Yue Sun[a], Deedee Kommers[b], Tao Tan[a], Wenjin Wang[c], Xi Long[c], Caifeng Shan[c], Carola van Pul[b], Ronald M. Aarts[a], Peter Andriessen[b], and Peter H.N. de With[a]

[a]Eindhoven University of Technology, 5612WH Eindhoven, The Netherlands
[b]Máxima Medical Center, 5504 DB Veldhoven, The Netherlands
[c]Philips Research, High Tech Campus 34, 5656AE Eindhoven, The Netherlands

ABSTRACT

Pain or discomfort exposure during hospitalization of preterm infants has an adverse effect on brain development. Contactless monitoring has been considered to be a promising approach for detecting infant pain and discomfort moments continuously. In this study, our main objective is to develop an automated discomfort detection system based on video monitoring, allowing caregivers to provide timely and appropriate treatments. The system first employs the optical flow to estimate infant body motion trajectories across video frames. Following the movement estimation, Log Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCCs) and Spectral Subband Centroid Frequency (SSCF) features are calculated from the One-Dimensional (1D) motion signal. These features enable the representation of the 1D motion signals by Two-Dimensional (2D) time-frequency representations of the distribution of signal energy. Finally, deep Convolutional Neural Networks (CNNs) are applied on the 2D images for the binary - comfort/discomfort classification. The performance of the model is assessed using leave-one-infant-out cross-validation. Our algorithm was evaluated on a dataset containing 183 video segments recorded from 11 infants during 17 heel prick events, which is a pain stimulus associated with a routine care procedure. Experimental results showed an area under the receiver operating characteristic curve of 0.985 and an accuracy of 94.2%, which offers a promising possibility to deploy the proposed system in clinical practice.

Keywords: Computer-Aided Diagnosis, Convolutional Neural Networks, Deep Learning, Discomfort Detection, Infant Discomfort

1. INTRODUCTION

Preterm birth is defined as the case when infants are born before 37 completed weeks of gestation. Global prevalence estimation of preterm birth is 9.6%, influencing approximately 12.9 million infants in 2005.[1] Pain/discomfort in infants has received considerable attention from researchers. Early pain experiences of preterm infants have long-term effects on their development, which include alterations in sensory processing and delay in neurological development. Cumulative pain-related discomfort can lead to long-term perseverance of central nervous system changes and similarly, long-term changes in responsiveness of the neuroendocrine and immune systems to stress at maturity.[2,3] In adults, self-reporting is regarded as the gold standard of pain assessment measurement among patients, since it provides the most reliable indication of pain.[4] However, preterm neonates are not capable of interpreting pain or discomfort in a manner similar to that of adults. Thus, monitoring is required in order to detect pain/discomfort immediately when infants start suffering, which allows caregivers to perform appropriate treatments.

Several pain/comfort-scales have been developed to assist healthcare professionals in assessing the pain or discomfort levels of an infant.[5,6] Each scale is scored by healthcare professionals after observing infants for several minutes. However, infants are only assessed a few times a day ("spot measurement") without continuous

Further author information:
Yue Sun: E-mail: y.sun1@tue.nl

Medical Imaging 2020: Computer-Aided Diagnosis, edited by Horst K. Hahn, Maciej A. Mazurowski, Proc. of SPIE Vol. 11314, 113144B · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2549250

monitoring, which may leave many discomfort moments unnoticed. In this regard, an automatic discomfort/pain-assessment method is needed. The objective of this paper is to develop an automated video-based discomfort detection system for infants.

In the past several years, there has been an increasing interest in understanding infant behavior using machine-learning based methods.[7] Emotion classification by analyzing facial expression has been investigated.[8,9] However, the faces of preterm infants in the Neonatal Intensive Care Unit (NICU) is often occluded by nasal cannulas, oxygen masks or feeding tubes. Thus, discomfort detection fully relying on a facial expression recognition method is not practical. Previously, we have proposed an alternative approach - assessing infant pain/discomfort based on body movement from videos.[10] Statistical and spectral features were extracted from the estimated body motion signal, which was followed by adopting a Support Vector Machine (SVM) classifier to differentiate discomfort status from comfort. However, these features are sensitive to parameter tuning of the SVM.

In this paper, we propose a deep learning-based scheme by first converting each motion signal segment into an image representation, and then applying state-of-the-art deep learning networks for the image classification task. Analyzing image representation of motion signals facilitates to use richer frequency-related information in addition to the motion time series. This approach results in the following contributions: 1) we define an algorithm for segmenting motion and then convert each segment to time-frequency image representations, 2) we employ a deep learning scheme by Convolutional Neural Networks (CNNs) to classify the feature-images, and 3) an automated video-based system for accurately detecting discomfort moments in preterm newborns in NICU is realized.

2. METHODS

2.1 Study design

In our work, Heel Prick (HP) is a well-known recurring pain stimulus, which was used as a stimulus to study infant response to pain. The study was conducted at the Máxima Medical Center in Veldhoven, the Netherlands. As the experimental procedure (HP) is part of regular neonatal care, the ethical committee of the Máxima Medical Center provided a waiver [N17.178] for this study. For all infants, written consent was obtained from the parents. A camera (uEye UI-222x, IDS imaging, Germany) was used to record the infant face and upper body in a fixed position.

Videos segments were manually annotated for comfort/discomfort by a medical doctor along the HP procedure. The comfort/discomfort video segments were labeled according to the timeline relative to the heel prick intervention by visual observation. Comfort video segments were annotated from the baseline before the prick and the period when each infant returned to baseline after the heel prick. Discomfort video segments were annotated from the starting point of heel prick to several minutes after the heel prick was finished. Each video segment is associated with only one infant state (comfort or discomfort).

Eleven infants with an average gestational age of 31 weeks (range 27^{+1}-38^{+5} weeks) were recorded. In total, we obtained 99 discomfort (2,738 seconds) and 84 comfort (3,429 seconds) video segments from 17 HP events.

2.2 Motion Estimation

We first employ an optical flow algorithm to estimate pixel-based motion vectors between adjacent video frames to extract body motion of the infants for each video segment. The utilized optical flow method proposed by Farnebäck *et al.*[11] enables capturing the motion from individual body parts. Dense optical flow is estimated by modeling the neighborhoods of each pixel using quadratic polynomials, and the optimization is done at pixel-cluster level rather than individual pixel level. We accumulate the magnitude values of all motion vectors for each video frame. Hence, all the summed magnitude values comprise a One-Dimensional (1D) signal vector of motion velocity magnitude for each video segment. We further estimate the motion acceleration rate by taking the first derivative of the velocity magnitude signal.

2.3 Image Representation

Following motion estimation, each 1D signal (motion acceleration rate) is clipped to 10-sec. long segments for further processing. One important step for the classification task is to identify the primary information, characterized by the discomfort motion pattern, while discarding other details that carry background noise, random movements, etc. The shape of the 1D signal manifests itself in the envelope of the short-time power spectrum. For this reason, we aim to extract features from the envelope shape to represent motion signals indicating the type of behavior. Three methods of feature extraction are investigated and compared for data representation, namely: Log Mel-spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), and Spectral Subband Centroid Frequency (SSCF). These methods are effective at extracting and combining the frequency and magnitude information from the power spectrum.[12–15]

MFCCs are features widely used to represent characteristics of signals in voice recognition.[16] We first execute overlapping sliding windows over the input signal segment, and then compute the Fourier transform over each window. A Mel-filterbank is further applied, and the energies within each filter are summed up. The Mel-spectrogram is therefore obtained as a graph of the energy content as function of Mel scales versus time windows. We take the logarithm (log) of the filterbank energies and employ the derived Log Mel-spectrogram as our first type of image representation (see Fig. 1(b) and Fig. 2(b)).

Following the Log Mel-spectrogram calculation, a Discrete Cosine Transform is applied on the log-filterbank energies, resulting in 12 MFCC values per sliding window. The total energy per sliding window is also included as a feature. As a result, 13 MFCC feature values are obtained in total. Concatenating these features leads to a time-frequency representation that can be visualized as a heat map, which is our second image representation - MFCCs. In total, each heat map consists of time windows represented on the horizontal axis, and 13 MFCC filterbanks represented on the vertical axis (see Fig. 1(c) and Fig. 2(c)).

Furthermore, we compute the SSCF from 1D signal segments. The SSCF represents the centroid frequency in each subband whereas in MFCC features, the power spectrum in a given subband is smoothed. Therefore, the SSCF provides supplementary information to MFCCs (see Fig. 1(d) and Fig. 2(d)).

2.4 Classification

The results of image representation processing allow each 10-second instance of the motion signal data to be processed as an image, so that energy values over time can be visualized as a heat map.

The derived heat maps are further classified using deep convolutional neural networks. A ResNet (See Fig. 3) is used as our classification model. He *et al.*[17]proposed ResNet, which is a network architecture where the layers are explicitly reformulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. It was shown that ResNet is easier to optimize, and can gain accuracy from considerably increased depth on the ILSVRC 2015 classification task. To alleviate the limitation of small sample size in this study, we employ transfer learning using the pretrained models. We use ResNet with 18 residual blocks, pretrained with the ImageNet dataset.

2.5 Evaluation

The proposed method is evaluated by performing leave-one-infant-out cross-validation to obtain an unbiased label for each video segment. The training set is further split into a real training set (70%) and a validation set (30%) on patient level. The validation set is used to refine the model from each training epoch. We perform training with the number of epochs = 25, a batch size of 16, Adam optimizer,[18] and without dropout.

The Receiver Operating Characteristic (ROC) is plotted to evaluate the performance with the value of the Area Under the ROC Curve (AUC). The classification accuracy and confusion matrix are also computed and reported as evaluation metrics.

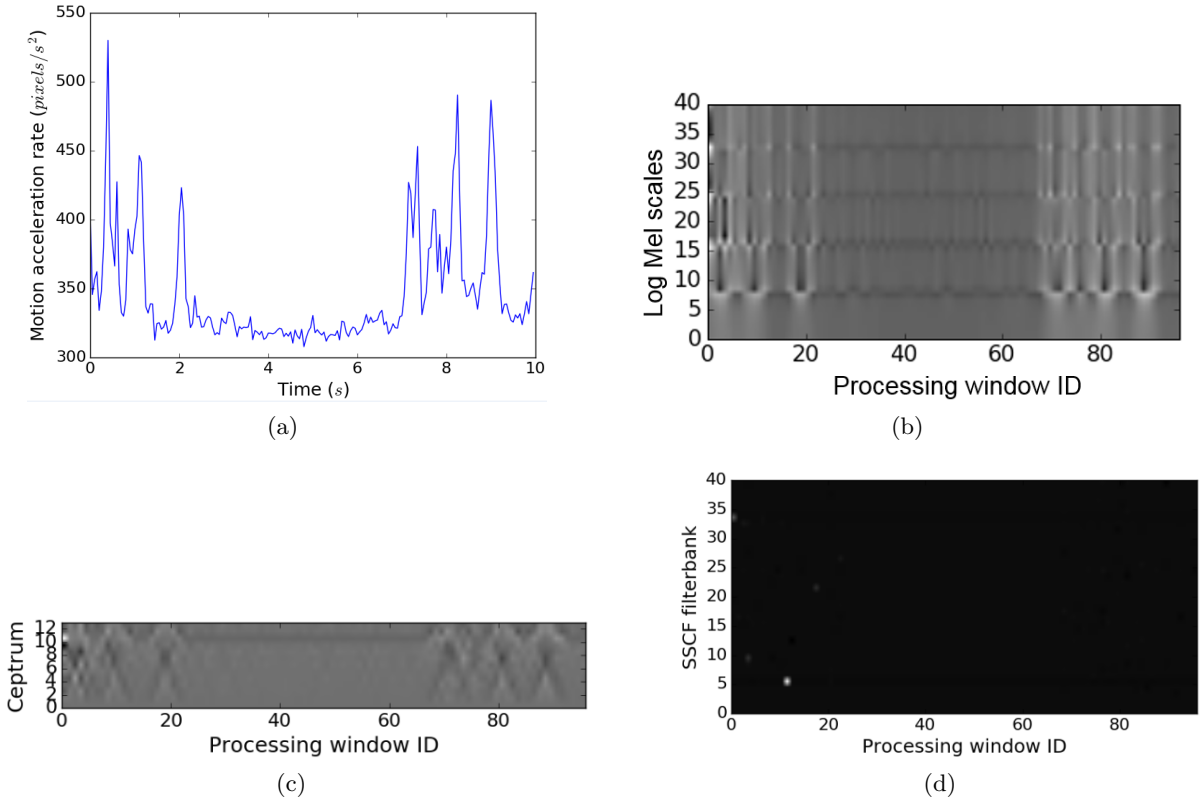

Figure 1. Example of a discomfort motion segment. (a) Extracted 1D motion signal, which is analyzed further with a sliding window (window size = 500 ms and step size = 100 ms). Feature-images for the 10-sec motion segment are shown in (b) Log Mel-spectrogram, (c) MFCCs and (d) SSCF, which all use the same window processing.

3. EXPERIMENTAL RESULTS

We conducted experiments with different frameworks of applying ResNet: 1) only fine-tuning the Fully Connected Layers (FCL) of a pretrained ResNet, 2) fine-tuning all layers of a pretrained ResNet, and 3) directly training ResNet using our dataset. Combining all the features together, the performance of applying different frameworks for the binary classification is shown in table 1. The results from our previous work[10] using handcrafted features on the same dataset are added in the table for reference. Fig. 4 shows the normalized confusion matrix for only fine-tuning the fully connected layers. Fig. 5 represents the corresponding ROC with the AUC of 0.985. The ROC curve indicates that approximately 90% comfort video segments can be correctly determined by our automated system without missing any discomfort moments.

The AUC values for applying each type of image representation individually are 0.978 for Log Mel-spectrogram, 0.961 for MFCCs and 0.677 for SSCF.

Table 1. Performance of different training schemes in terms of classification accuracy, and AUCs.

Training scheme	Accuracy	AUC
Fine-tuning only FCL	94.2%	0.985
Fine-tuning all layers	93.3%	0.978
Training from scratch	80.8%	0.878
Handcrafted features[10]	86.0%	0.940

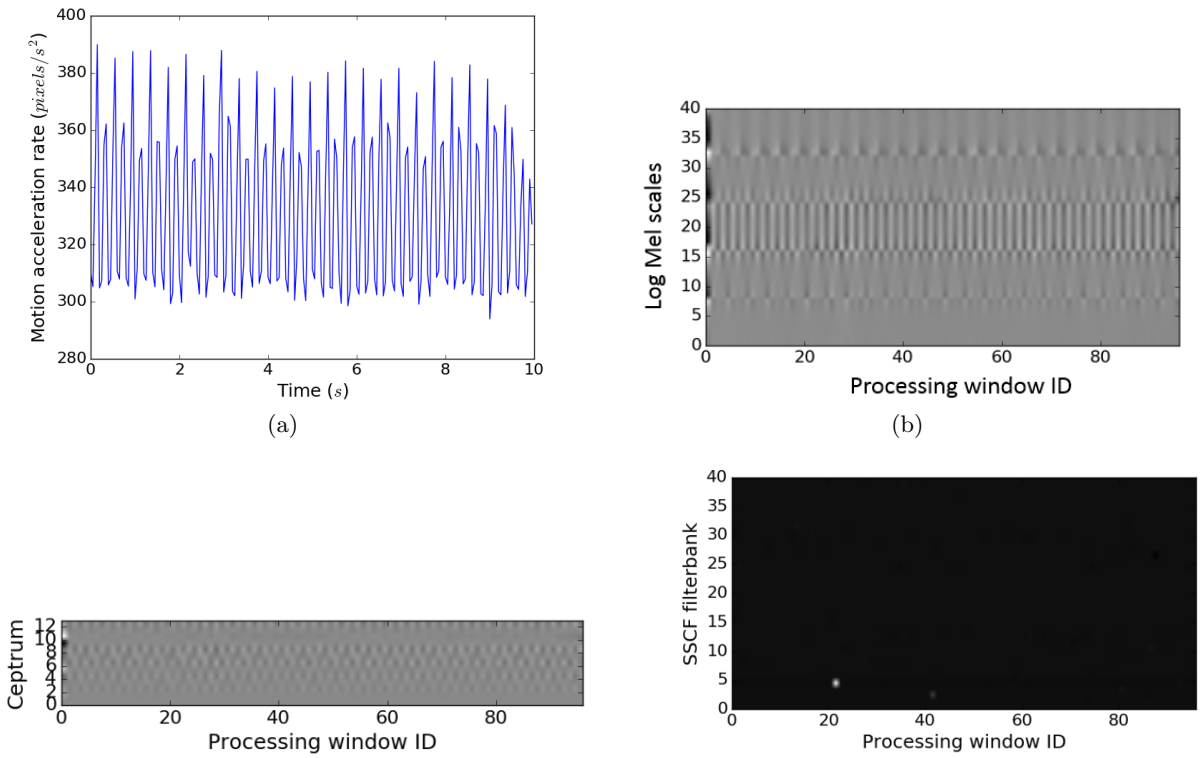

Figure 2. Example of a comfort motion segment. (a) Extracted 1D motion signal, which is analyzed further with a sliding window (window size = 500 ms and step size = 100 ms). Feature-images for the 10-sec motion segment are shown in (b) Log Mel-spectrogram, (c) MFCCs and (d) SSCF, which all use the same window processing.

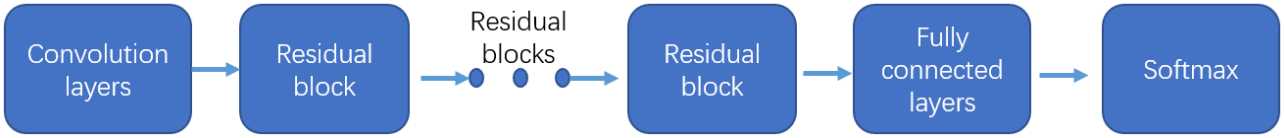

Figure 3. The general architecture of the ResNet.

4. BREAKTHROUGH WORK AND CONCLUSION

In this work, we propose an automated video-based system using deep learning that can differentiate discomfort status of infants from comfort status by analyzing motion patterns. When infants experience discomfort moments, the system allows drawing attention of caregivers to them. The processing chain includes three steps: 1) 1D signal extraction using optical flow, 2) converting the 1D signal to feature-image representation, and 3) deep learning classification of the feature-images. For each video segment, three image representations characterizing motion trajectories are extracted from the motion signals in the time and frequency domain. An AUC of 0.985 is achieved, which is promising for use in clinical practice. The highest AUC is achieved when combining all three image representations by transfer learning from a pretrained ResNet, which proves that the different image representations are all contributing and complementary. In the future, more infant data will be collected and used for evaluating our system.

ACKNOWLEDGMENTS

The authors would like to thank all the medical staff from the NICU department of Máxima Medical Center, Veldhoven for the cooperation and assistance in collecting the infant videos.

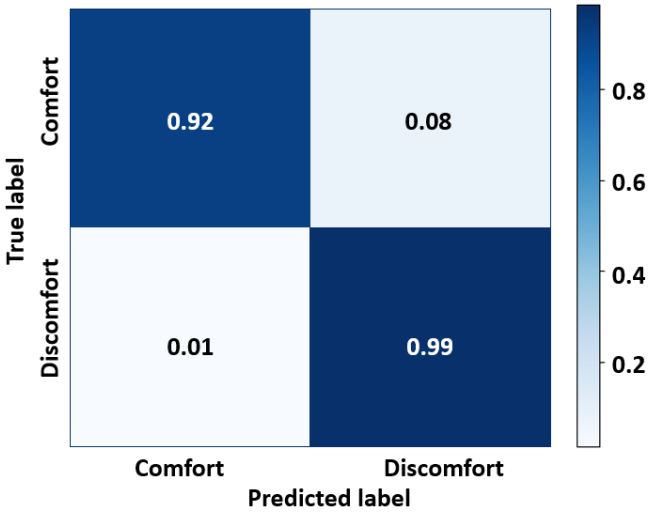

Figure 4. Normalized confusion matrix of comfort and discomfort classification with only fine-tuning the fully connected layers of ResNet.

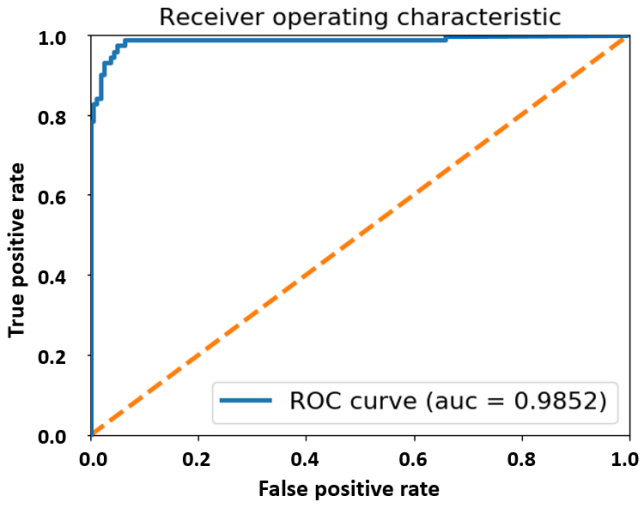

Figure 5. ROC curve of binary classification of comfort and discomfort with only fine-tuning the fully connected layers of a pretrained ResNet.

REFERENCES

[1] Beck, S., Wojdyla, D., Say, L., Betran, A. P., Merialdi, M., Requejo, J. H., Rubens, C., Menon, R., and Van Look, P. F., "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity," *Bulletin of the World Health Organization* **88**, 31–38 (2010).

[2] Page, G. G., "Are there long-term consequences of pain in newborn or very young infants?," *The Journal of perinatal education* **13**(3), 10 (2004).

[3] Petrini, J. R., Dias, T., McCormick, M. C., Massolo, M. L., Green, N. S., and Escobar, G. J., "Increased risk of adverse neurological development for late preterm infants," *The Journal of Pediatrics* **154**(2), 169–176 (2009).

[4] Melzack, R. and Katz, J., "The gate control theory: Reaching for the brain," *Pain: psychological perspectives* , 13–34 (2004).

[5] Ambuel, B., Hamlett, K. W., Marx, C. M., and Blumer, J. L., "Assessing distress in pediatric intensive care environments: the comfort scale," *Journal of pediatric psychology* **17**(1), 95–109 (1992).

[6] Norden, J., Hannallah, R., Getson, P., O'Donnell, R., Kelliher, G., and Walker, N., "Reliability of an objective pain scale in children," *Journal of pain and symptom management* **6**(3), 196 (1991).

[7] Zamzmi, G., Kasturi, R., Goldgof, D., Zhi, R., Ashmeade, T., and Sun, Y., "A review of automated pain assessment in infants: Features, classification tasks, and databases," *IEEE reviews in biomedical engineering* **11**, 77–96 (2018).

[8] Kotsia, I. and Pitas, I., "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing* **16**(1), 172–187 (2007).

[9] Sun, Y., Shan, C., Tan, T., Long, X., Pourtaherian, A., Zinger, S., et al., "Video-based discomfort detection for infants," *Machine Vision and Applications* , 1–12 (2018).

[10] Sun, Y., Kommers, D., Wang, W., Joshi, R., Shan, C., Tan, T., Aarts, R. M., van Pul, C., Andriessen, P., and de With, P. H., "Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis," in [*2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*], 5995–5999, IEEE (2019).

[11] Farnebäck, G., "Two-frame motion estimation based on polynomial expansion," in [*Scandinavian Conference on Image Analysis*], 363–370, Springer (2003).

[12] Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al., "Multiple classifier systems for the classification of audio-visual emotional states," in [*International Conference on Affective Computing and Intelligent Interaction*], 359–368, Springer (2011).

[13] Bilik, I. and Khomchuk, P., "Minimum divergence approaches for robust classification of ground moving targets," *IEEE Transactions on Aerospace and Electronic Systems* **48**(1), 581–603 (2012).

[14] Salamon, J. and Bello, J. P., "Feature learning with deep scattering for urban sound analysis," in [*2015 23rd European Signal Processing Conference (EUSIPCO)*], 724–728, IEEE (2015).

[15] Quiceno-Manrique, A., Alonso-Hernandez, J., Travieso-Gonzalez, C., Ferrer-Ballester, M., and Castellanos-Dominguez, G., "Detection of obstructive sleep apnea in ecg recordings using time-frequency distributions and dynamic features," in [*2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*], 5559–5562, IEEE (2009).

[16] Logan, B. et al., "Mel frequency cepstral coefficients for music modeling.," in [*ISMIR*], **270**, 1–11 (2000).

[17] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[18] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *International Conference on Learning Representations* (2015).