Detecting discomfort in infants through facial expressions

Detecting Discomfort in Infants through Facial Expressions

YUE SUN, Eindhoven University of Technology, The Netherlands
CAIFENG SHAN*, Philips Research, High Tech Campus 34, The Netherlands
TAO TAN, Eindhoven University of Technology, The Netherlands
TONG TONG, Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts, General Hospital, Harvard Medical School, USA
WENJIN WANG, Philips Research, High Tech Campus 34, The Netherlands
ARASH POURTAHERIAN, Eindhoven University of Technology, The Netherlands
PETER H.N. DE WITH, Eindhoven University of Technology, The Netherlands

Detecting discomfort status of infants is particularly clinically relevant. Late treatment on discomfort infants can lead to adverse problems such as abnormal brain development, central nervous system damage and changes in responsiveness of the neuroendocrine and immune systems to stress at maturity. In this study, we exploit deep Convolutional Neural Network (CNN) algorithms to address the problem of discomfort detection for infants by analyzing their facial expressions. A dataset of 55 videos about facial expressions, recorded from 24 infants, is used in our study. Given the limited available data for training, we employ a pre-trained CNN model, which is followed by fine-tuning the networks using a public dataset with labeled facial expressions (the Shoulder-Pain dataset). The CNNs are further refined with our data of infants. Using a two-fold cross-validation, we achieve an Area Under the Curve (AUC) value of 0.96, which is substantially higher than the results without any pre-training steps (AUC=0.77). Our method also achieves better results than the existing method based on handcrafted features. By fusing individual frame results, the AUC is further improved from 0.96 to 0.98. The proposed system has great potential for continuous discomfort and pain monitoring in clinical practice.

Additional Key Words and Phrases: Discomfort detection, Facial expression recognition, Deep learning, DenseNet, Fine-tuning, Transfer learning

Reference Format:
Yue Sun, Caifeng Shan, Tao Tan, Tong Tong, Wenjin Wang, Arash Pourtaherian, and Peter H.N. de With. 2019. Detecting Discomfort in Infants through Facial Expressions. 1, 1 (November 2019), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Discomfort and pain assessment of infants is important, not only for their wellbeing but also for their brain development. Infants are particularly vulnerable to pain-related stress, due to their low pain threshold, sensitization from repeated pain, and immature systems for maintaining homeostasis [6][7]. Frequent discomfort or pain in infants undergoing prolonged hospitalization can cause complications, such as delay in cognitive and motor development [9]. Cumulative discomfort or pain is also associated with altered brain development and, consequently, deficits in neurological outcomes [28][45][2][31]. Therefore, continuous discomfort (or pain) monitoring of infants is needed to prevent the complications mentioned above as much as possible. The monitoring can help caregivers understand the severity of the infant status and develop appropriate treatments.

At present, discomfort and pain monitoring for infants is performed manually by healthcare professionals, who check the behavioral (e.g., facial expression, body movement and crying sound) and/or physiological (e.g., vital signs) indicators of infants. This discomfort and pain assessment is of high cost, time-consuming and subjective in assessment [32][3]. Furthermore, infants are observed only a few times a day during short intervals ("spot measurements") without continuous monitoring, which likely leaves many discomfort moments unnoticed. The intermittent assessment might lead to under/misdiagnosis and therefore delayed/incorrect treatment.

*Corresponding author. Email: caifeng.shan@philips.com

Authors' addresses: Yue Sun, Eindhoven University of Technology, Eindhoven, 5612 WH, The Netherlands; Caifeng Shan, Philips Research, High Tech Campus 34, Eindhoven, 5656 AE, The Netherlands, caifeng.shan@philips.com; Tao Tan, Eindhoven University of Technology, Eindhoven, 5612 WH, The Netherlands, t.tan1@tue.nl; Tong Tong, Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts, General Hospital, Harvard Medical School, Boston, MA 02129, USA; Wenjin Wang, Philips Research, High Tech Campus 34, Eindhoven, 5656 AE, The Netherlands; Arash Pourtaherian, Eindhoven University of Technology, Eindhoven, 5612 WH, The Netherlands; Peter H.N. de With, Eindhoven University of Technology, Eindhoven, 5612 WH, The Netherlands.

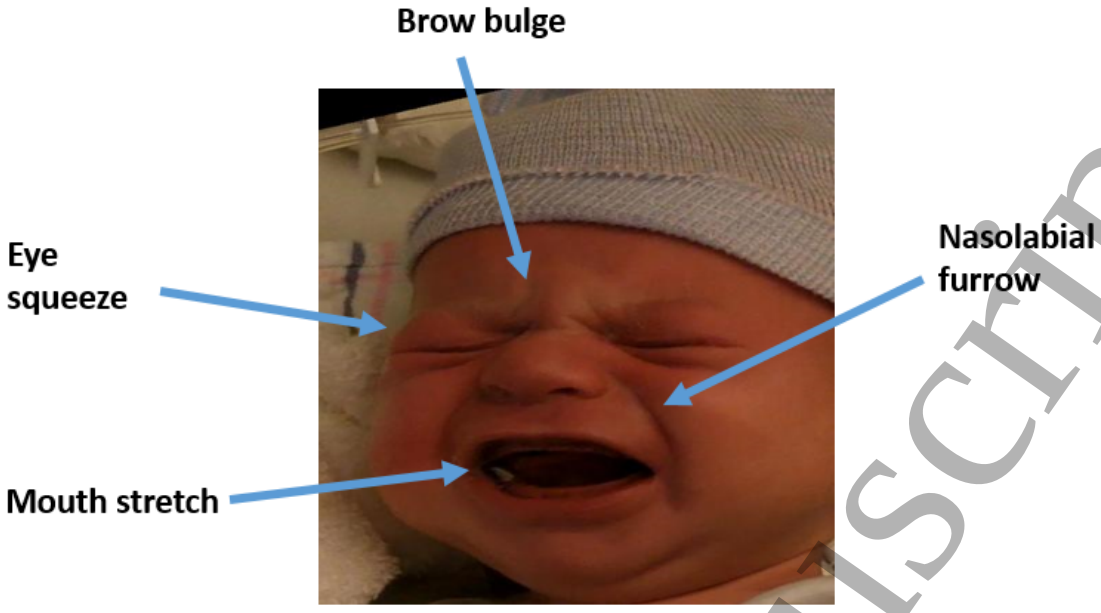


Fig. 1. Primal face of pain.

In this work, we propose an automatic discomfort detection method for infants by analyzing their facial expressions with deep Convolutional Neural Networks (CNN) algorithms. Facial expression appears to be one of the behavior indicators of discomfort or pain. The Primal Face of Pain (PFP), as shown in Figure 1, is an instinctive and universal facial expression associated with pain [35]. Automatic detection of discomfort (or pain) in videos by analyzing the facial expression of infants is a potential solution for continuous monitoring. It is difficult to collect facial expression video data of infants due to privacy and ethical issues. We have managed to collect a dataset of 55 videos from 24 infants. However, to train a deep CNN algorithm, a large dataset is required for a good generalization ability of the network, and a small dataset easily leads to overfitting. To address this, we utilize transfer learning by employing several training steps based on various sources. More specifically, we adopt a pre-trained model, which is followed by training the networks using a public dataset [25] with labeled facial expressions for pain assessment (first fine-tuning). The networks are then further fine-tuned with our dataset of infants (second fine-tuning). The entire workflow is depicted in Figure 2. The proposed method has been applied for infant-independent discomfort detection on our dataset. Using a two-fold cross-validation, we achieve an Area Under the Curve (AUC) value of 0.96, which is substantially higher than the results without any pre-training steps (AUC=0.77). Compared to the existing method using handcrafted features [42], the AUC of our method also increases 10% on the same dataset.

The remainder of this paper is organized as follows. In Section 2, related work on pain/discomfort detection and CNN-based classification is described. Section 3 elaborates our method, and Section 4 explains the experimental results. Finally, Section 5 discusses the results and draws conclusions.

2 RELATED WORK

There are very limited studies on automatic pain or discomfort detection for infants in videos [40][8][42][46]. Sikka *et al.* [40] proposed a Facial Action Coding System (FACS) and associated methods to describe the facial expressions of pain of children. Fotiadou *et al.* [8] presented a discomfort detection system utilizing the active appearance model (AAM) and a Support Vector Machines (SVM) classifier. The system achieved an AUC of 0.98 by evaluating in 15 videos from 8 infants. However, for each baby, a set of landmarks need to be placed manually for initializing the AAM mesh. Sun *et al.* [42] designed appearance and geometric features to describe infant faces for discomfort detection, and achieved an AUC of 0.87. However, the features were handcrafted, and SVM was used as the classifier.

In the past, significant attention was paid to facial expression recognition of adults [38][33]. Kotsia and Pitas [20] proposed two methods for facial expression recognition: 1) estimating geometrical displacement of certain selected Candide grid nodes, which was followed by a multiclass SVM system, and 2) a Facial Action Units (FAUs) based approach. The recognition accuracy of 99.7% and 95.1% was achieved when using the multiclass SVM and FAU-based method, respectively. Shan *et al.* [39] presented a comprehensive empirical study of facial expression recognition based on Local Binary Patterns (LBP) and illustrated that

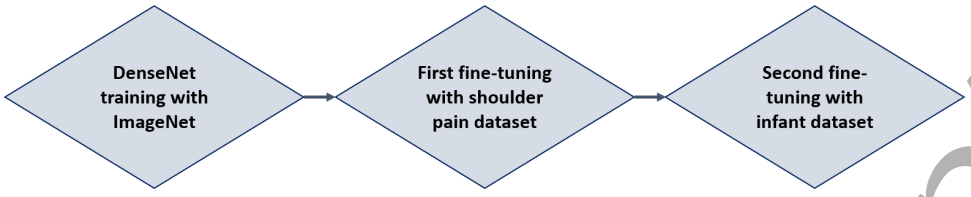


Fig. 2. Workflow of the proposed discomfort detection method using CNNs with fine-tuning steps.

LBP features performed stably and robustly over a useful range of low resolutions of face images. Different machine learning methods, such as Adaboost [37], have been further exploited for facial expression classification. Neshov and Manolova [26] used supervised descent method and scale invariant feature transform, which yielded a very good recognition rate (more than 95.7%). A hierarchical unsupervised feature learning approach was employed by Kharghanian *et al.* [17] to extract the features detecting pain from facial images based on a convolutional deep belief network. The AUC of the Receiver Operating Characteristic (ROC) was near 95%. Lucey *et al.* [24] showed that the AAM-based system can overcome the facial deformation and head motion and yield significant improvements in both the FAU and pain detection. Ashraf *et al.* [1] explored various face representations derived from AAMs for detecting pain from faces, and demonstrated that decoupling a face into separate non-rigid shape and appearance components offers significant performance improvement. Hammal *et al.* [10] applied a set of Log-Normal filters consisting of 7 frequencies and 15 orientations to extract 9,216 features for pain estimation and the statistical analysis - F1 measure (the harmonic mean of the precision and recall) - for each level of pain intensity ranged from 91% to 96%. Littlewort *et al.* [22] proposed an automated facial expression recognition system to differentiate real pain from fake pain. A 20-channel output stream of facial action detectors from the FACS was passed to a classification stage to determine the labels. An accuracy of 88% was obtained for the subject-independent discrimination of real versus fake pain. Hazelhoff *et al.* [12] developed a prototype of an automated video survey system by analyzing facial expression. The method first localized eye, eyebrow and mouth regions, which was followed by employing a hierarchical classifier to discriminate between different behavioral states of sleep, awake and cry. An accuracy of 95% was achieved. A similar method [11] was applied on a dataset of Neonatal Intensive Care Unit (NICU), which resulted in an accuracy of 88%. Zhao *et al.* [47] proposed a novel Set-to-Set (S2S) distance measure to calculate the similarity between two sets with the aim to improve the recognition accuracy for faces with real-world challenges, as compared to traditional feature-average pooling and score-average pooling. However, this method still depends on the effectiveness of the extracted features. Ding *et al.* [4] proposed a deep confidence network (DECODE) for robust training. The method adopted an effective confidence evaluation module, which assigned small training weights to suspicious samples to suppress the influence of noisy data. The weighted training data was also used to update the weight values after each iteration. Evaluation of this method was carried out on several datasets, where the effectiveness of DECODE was demonstrated.

Recently, CNN has become a powerful tool for automated two- and three-dimensional image classification. Raghuvanshi *et al.* [30] classified images of human faces into discrete emotion categories using CNNs and experimented with different architectures and methods, such as fractional max-pooling and fine-tuning, ultimately achieving an accuracy of 48% in a seven-class classification task. Lopes [23] proposed a simple solution for facial expression recognition that uses a combination of CNN and a specific image preprocessing step for extracting only expression-specific features from a face image. The proposed method achieved competitive results when compared with other facial expression recognition methods. Wang *et al.* [44] proposed a network that fine-tuned a state-of-the-art face verification CNN network using a regularized regression loss and additional data with expression labels, which achieved the state-of-the-art performance.

3 METHODOLOGY

3.1 Preprocessing

To be able to classify facial expression, we first need to extract the face area from videos. The face Region of Interest (ROI) is generated based on the workflow of face detection and normalization. The selected ROI is passed on to the next step as the input to the discomfort/comfort classification. Given an input video frame, 68 facial landmarks are first localized using dlib face landmark detector [18], which is an implementation of Kazemi *et al.* [16]. The 68 landmarks include points on the face such as the corners of the eyes, mouth, along with the eyebrows, along with the boundary of the face.

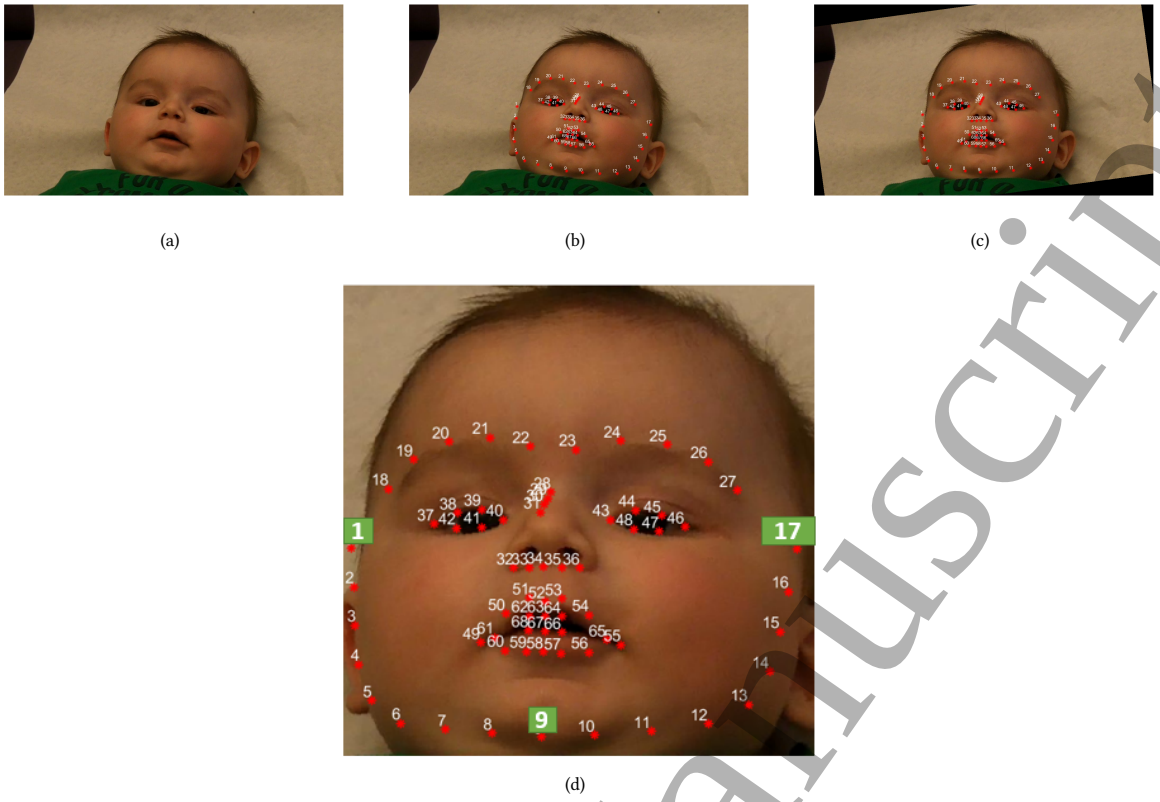


Fig. 3. (a) Sample original face image, (b) the original face image with identified 68 facial landmarks, (c) in-plane head rotation corrected, and (d) the final normalized face ROI with the corresponding 68 landmarks, of which landmark 1, 9 and 17 are highlighted.

Once the 68 landmarks are identified, the middle point between two inner eye corner points is used as a reference point to rotate the image. We rotate the image to the position where the line connecting the two eye corners points is horizontal. Thus, this step corrects the in-plane rotation variance of the face. We select Landmark 1 as the leftmost point, Landmark 17 as the rightmost, and Landmark 9 as the bottom-most points to define the left, right, and bottom boundaries of the face ROI. For the top boundary, the horizontal line that has the equal vertical distance from the inner eye center as Landmark 9 is chosen. A margin of 20 pixels is added to all the boundaries to cover the whole face and avoid losing facial information. Finally, all face images are cropped and then resized to 224×224 pixels using bilinear interpolation in order to suit the required input image size for CNNs. Figure 3 exemplify an original face image with a corresponding normalized face ROI and the detected 68 facial landmarks.

3.2 CNN model

Inspired by the performance of CNNs [13, 15, 21, 41], we propose to use a CNN model as a solution for our problem. The CNN usually contains several pairs of a convolutional layer and a max-pooling layer for non-linear down-sampling, followed by fully connected layers. Compared to the traditional facial expression systems based on handcrafted features [42], CNN has the advantage of self-learning. The parameters in CNN are automatically learned from the data and researchers do not need to design complicated features as input.

In this work, the adopted base model is DenseNet [14], which is a network architecture where each layer is directly connected to every other layer in a feed-forward fashion within each dense block. For each layer, the feature maps of all preceding layers are treated as separate inputs, whereas its own feature maps are passed on as inputs to all subsequent layers. DenseNet has the advantage of alleviating the vanishing-gradient problem, strengthening feature propagation, and reducing the number of parameters. In the referred work, this connectivity pattern yields the state-of-the-art classification performance on CIFAR10/100 (with or without data augmentation) and SVHN [14]. On the large-scale ILSVRC 2012 (ImageNet) dataset, DenseNet achieves a similar performance as ResNet, while using less than half of the number of parameters and roughly half of the number of floating-point operations (Flops). Figure 4 shows the CNN architecture of our work. The number 121

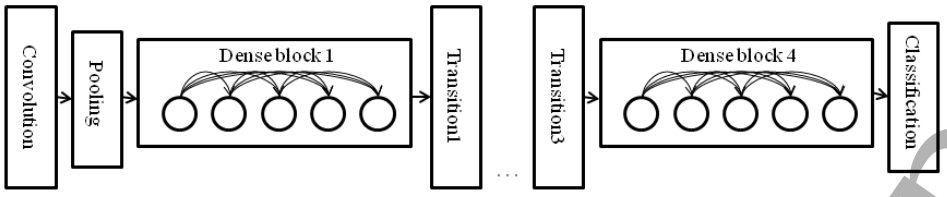


Fig. 4. Structure of the DenseNet 121 networks used in our system.

corresponds to the number of layers with trainable weights (excluded batch normalization layer), i.e. convolutional layers and fully connected layers. The additional 5 layers include the initial 7×7 convolutional layer, 3 transitional layers, and a fully connected layer. Between network blocks, the processing steps are convolution and pooling, which consist of a batch normalization layer and an 1×1 convolutional layer followed by a 2×2 average pooling layer. The Dense block is identical to the one introduced in [14], except for the last output layer, which is modified according to the task of our study because the number of classes is different. This specific CNN model is chosen due to its computation efficiency for the goal of a real-time, or semi-real-time application.

3.3 Transfer learning

In our work, a significant challenge is that a very limited number of infant videos are available. This is a common problem in the medical field, due to privacy issues and the boundaries on expenses for medical equipment. One effective solution is to use transfer learning [27][43], to address the problem of limited availability of labeled data. For the purpose of transfer learning, we can preserve all pre-trained layers prior to the last output layer and connect these layers to a new layer for the new classification problem. To train the networks for the new dataset, we allow the parameters from the fully connected layers of the network to be updated or optimized. The other choice is to fine-tune more layers, or even the whole set of pre-trained network layers. It is also possible to keep the first convolutional layer fixed, as this layer is often used for edge extraction [36], which is common for generic image processing problems. Therefore, as discussed earlier, since not all parameters are re-trained or trained from scratch, the transfer learning is beneficial to the problems with a small labeled dataset. In this work, to fully exploit the learning power of CNNs, we have chosen to fine-tune all parameters of the pre-trained DenseNet model. First, we obtain a DenseNet model trained on ImageNet data [14]. The first fine-tuning is based on a public dataset, which is followed by the second-tuning of using our own data of infants. The ImageNet dataset is also first resampled to 256×256 pixels using bilinear interpolation and then the patch of 224×224 in the center is cropped.

First fine-tuning step. Considering the limited number of samples in our dataset, instead of directly fine-tuning the DenseNet model using our own data, we first fine-tune the DenseNet model on a public dataset: the Shoulder-Pain dataset [25]. This dataset contains 200 videos of 25 subjects (48,398 frames in total) and is widely used for benchmarking the pain intensity estimation. For each frame, discrete pain intensities (0-15) according to Prkachin and Solomon [29] are provided by the database creators. In the same way as previous work [44][49][34][52], we quantify the original pain intensities within the range of [0, 15] to be in the range of [0, 5] for the purpose of data balancing. The pain intensities are discretized into 6 pain levels as follows: 0 (none), 1 (mild), 2 (discomforting), 3 (distressing), 4-5 (intense), and 6-15 (excruciating). The data balancing is performed in order to avoid overfitting of the test methods on the majority classes. Figure 5 shows 6 levels of the facial expression (status) from a woman in the Shoulder-Pain dataset. In the pre-trained DenseNet 121 model, we replace the original 1000 output nodes by 6 output nodes to represent 6 classes. The data is randomly split into a training dataset (70%) and validation dataset (30%). The validation dataset is used to select the best classifier, where the loss function achieves the minimum. To obtain the final label or class of each sample, we assign the label to the corresponding node in the last layer that gives the highest likelihood value. For training, we limit the number of epochs (m) to 50. Furthermore, in order to augment the number of training samples, images are randomly flipped horizontally, rotated within −10 degrees to +10 degrees, and translated within a distance of 20 pixels. Finally, we use the Adam algorithm [19] to optimize our loss function.

Second fine-tuning step. After the first fine-tuning step, we have a reasonable model for facial expression recognition. In this step, continuing with the previous model, we replace the 6 output nodes with 2 nodes, in order to match the classes to our problem statement: automated detection of discomfort and comfort. The parameters for the networks, training, and

Fig. 5. Facial expressions of six pain levels from a woman in the Shoulder-Pain dataset. For subfigures (a) through (f), the pain intensity score increases from 0 to 5.

data augmentation are the same as the previous step (see Figure 6 for the examples of the dataset, and Figure 7 for the corresponding post-preprocessing images). In order to obtain an unbiased evaluation of the classification performance, a two-fold cross-validation is employed to evaluate the performance of our method. Specifically, the input dataset is randomly divided into two equal parts at the patient level, where one part is left out for testing, and the other part is split again for training (70%) and validation (30%), to avoid bias. All the parameters are updated during validation. The classifier with the lowest loss based on the validation set is chosen as the best classifier and is used for testing. Such a procedure is repeated two times with a different part used for testing. We have pooled and evaluated the results from both parts and obtain the performance measurements.

4 EXPERIMENTAL RESULTS

4.1 Captured video materials

The study was conducted with videos recorded at the Máxima Medical Center (MMC) in Veldhoven, The Netherlands, by a hand-held high-definition camera (Xacti VPC-FH1BK). For all infants in the database, written consent was obtained from at least one of the parents. Data from 24 infants were collected in total. The faces were recorded when they were experiencing stressful moments including clinical treatment of heel prick, placing an intravenous (IV) line, venipuncture, vaccination, post-operative pain, and discomfort moments of the diaper change, feeling hungry or crying for attention. For 10 out of the 24 infants, the relaxed comfort state of resting or sleeping was also recorded. For 4 infants only their comfort moments are recorded. Thus, the image frames contain 1 to 2 emotions per subject. The number of infants regarding the recorded status of comfort/discomfort is summarized in Table 1. The duration of the video segments varies from less than 1 minute to several minutes.

Table 1. Dataset summarization.

Infant status	No. of videos
Comfort only	4
Discomfort only	10
Exhibiting both	10

The age of the 24 recorded infants ranges between 2 days and 13 months old. Three of the infants were born premature, and under 37 weeks at the time of recording. Examples of video frames in the dataset are shown in Figure 6. The resolution of each video frame is 1920×1080 pixels, and the frame rate is 30 frames per second (fps). The videos were recorded under uncontrolled lighting conditions. The labels of comfort/discomfort for each frame are annotated according to the consensus of 2 clinical experts. We have extracted video segments of which the infants are in supine position. Finally, a total of 16,837

Fig. 6. Examples of frames in the database of this study, where comfort frames are highlighted by green boxes, while discomfort cases are by red. Top four rows show the ten infants having both comfort and discomfort moments recorded. The 2nd and 4th rows are the comfort frames. The 1st and 3rd rows are for discomfort frames. The ten infants with only discomfort moments recorded are shown in the 5th and 6th rows. The four pictures in the bottom row are the infants with only comfort moments recorded. Three premature infants are outlined using yellow dotted rectangles.

frames are obtained, on which facial landmarks are detected. From all of the frames, 6,534 present comfort, and the rest 10,303 are discomfort frames. There are more discomfort samples in our dataset than comfort samples, since data collection in the hospital has focused on recording the discomfort moments of the infants.

4.2 Results without any pre-training

To show the effectiveness of pre-training with transfer learning, we have first performed the experiments by directly using our data to train a DenseNet classifier from scratch. The data augmentation procedure for the training data is the same as for training from scratch, the first and the second fine-tuning step. Results without any pre-trainings are directly trained from the model from scratch without using any ImageNet data. A weighted loss function was used according to the size of each class in the training data, in order to account for the nature of data imbalance. Given training samples from the entire 24 infants, a two-fold cross-validation is employed for evaluation. The two-fold cross-validation is the same as that described in Section 3.3 Transfer learning - Second fine-tuning step.

Figure 8 shows the normalized confusion matrix, based on the cross-validation without any pre-training. The obtained accuracy of all validated images is 52.4% and the accuracy values for the two classes of comfort and discomfort are 87% and 30%, respectively. Since in our application, the main focus is to detect discomfort moments of infants in time, the low detection rate for discomfort moments is not acceptable.

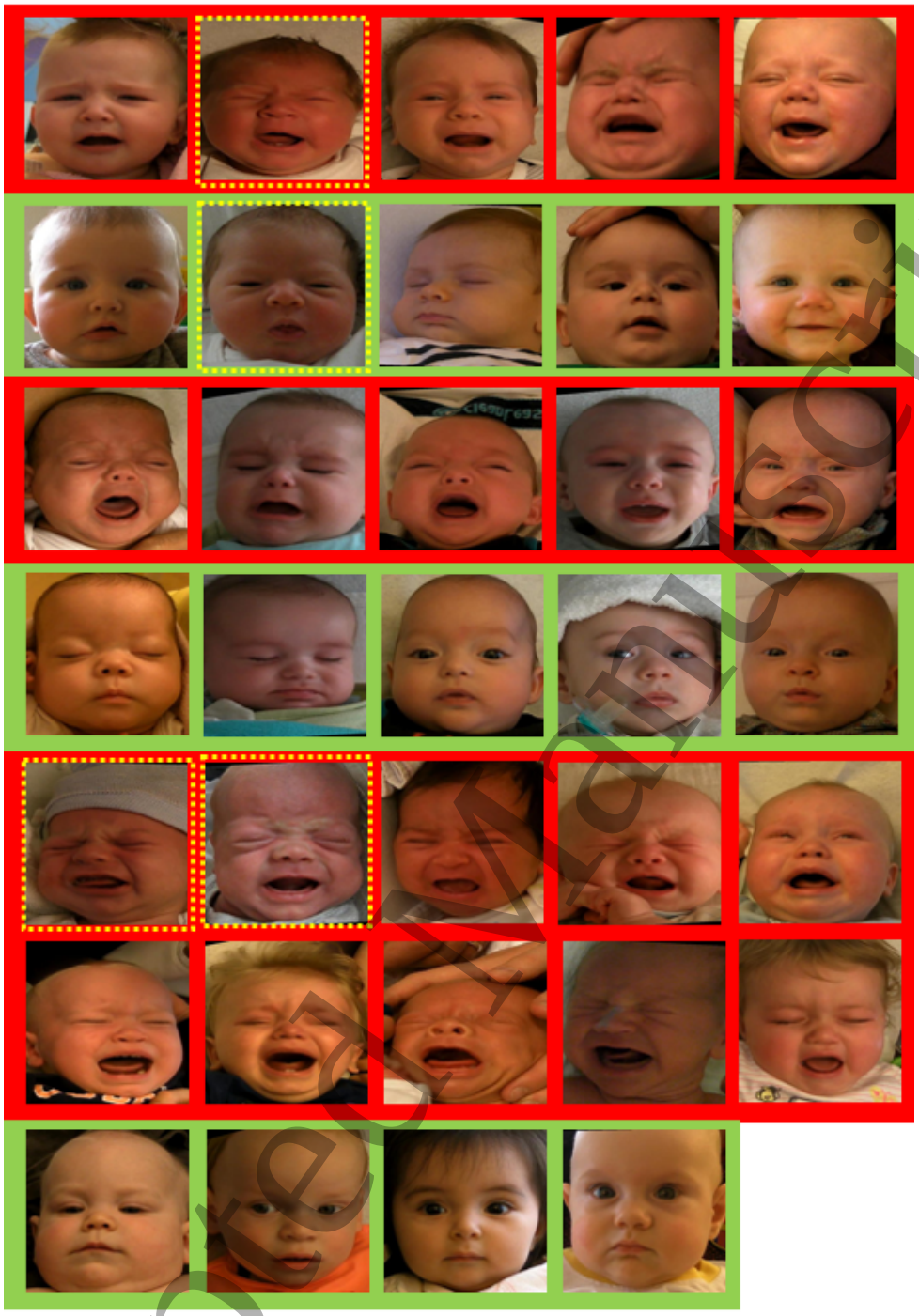


Fig. 7. Post-preprocessing (pre-networks) infant images of the frames in Figure 6 shown in the order corresponding to Figure 6. Comfort frames are highlighted by green boxes, while discomfort cases are by red. Three premature infants are outlined using yellow dotted rectangles.

4.3 Results from the first fine-tuning step

For the first fine-tuning step, the Shoulder-Pain dataset is used. The Shoulder-Pain dataset has been randomly split into a training dataset (70%) and validation dataset (30%). The classification accuracies are calculated based on the performance on the validation dataset, which shows that the overall accuracy is 85% and the accuracy values for the six classes of facial expressions are 86%, 81%, 79%, 78%, 86%, and 98%, respectively. We compared the performance of our model with existing methods [48][51] on the Shoulder-Pain dataset by calculating the Mean Absolute Error (MAE) as deviation from the ground-truth labels, Mean Squared Error (MSE) and the Pearson Correlation Coefficient (PCC) for the 6-level pain intensity classification (See Table 2).

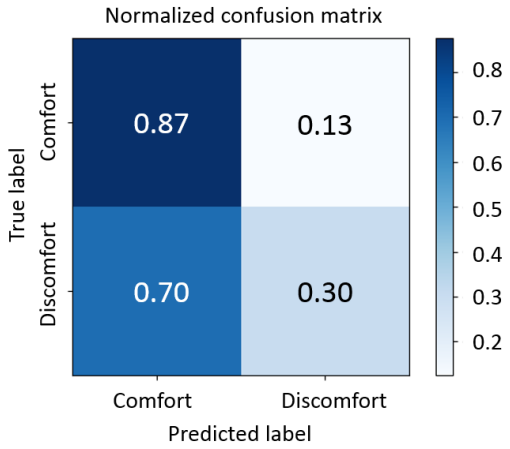


Fig. 8. Normalized confusion matrix of the model without any pre-training.

Comparing to the two existing methods, our method performs best by achieving the lowest MAE of 0.451 and highest PCC of 0.643.

Table 2. Performance comparison on shoulder-pain dataset.

Method	MAE	MSE	PCC
Our transfer-learning model	0.451	0.950	0.643
Ordinal information based regression [48]	1.025	N/A	0.600
Recurrent convolutional neural network regression [51]	0.810	N/A	0.601

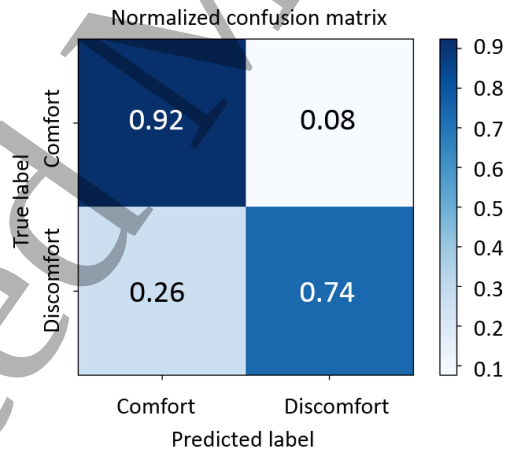


Fig. 9. Normalized confusion matrix of the method, with DenseNet pre-trained on ImageNet data and using the 2nd fine-tuning step on our data (i.e., without the 1st fine-tuning step).

4.4 Results from the second fine-tuning step

Figure 9 shows the confusion matrix and the normalized confusion matrix of the method when applying only the second fine-tuning step on our data based on cross-validation (i.e., without the first fine-tuning step). The accuracy of all validation images is 81% and the accuracy of the two classes of infant status are 92% and 74%. In comparison, Figure 10 portrays the normalized confusion matrix of the method including the two fine-tuning steps on our data based on cross-validation. The accuracy of all validation images is 91% and the accuracy of the two classes of infant status are 90% and 92%, respectively. The average accuracy of the three premature infants in the dataset is 85%, which shows that our system is also interesting to be considered for premature infant discomfort detection.

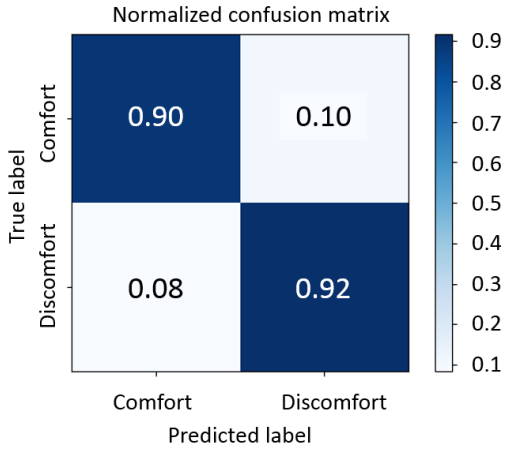


Fig. 10. Normalized confusion matrix of the proposed method with two fine-tuning steps on our data.

(a) (b)

Fig. 11. Examples of discomfort expressions being misclassified. (a) Discomfort status is misclassified as comfort status. (b) Comfort case is misclassified as discomfort.

We have also computed the ROC curve on our data without and with the first fine-tuning step (see Figure 12). AUC increases from 0.93 to 0.96. From the ROC curve, we can see that 78% of the normal status can be safely eliminated by our system, while the sensitivity preserved at a very high level.

Table 3. Measured classification performance including the classification accuracy (ACC) of all cases (AC), classification accuracy of comfort cases (CC), classification accuracy of discomfort cases (DC) and the AUCs with corresponding 95% confidence intervals (CIs) of different training schemes/methods.

Training method	ACC on AC	ACC on CC	ACC on DC	AUC	95% CI of AUC
Without any pre-training	52%	87%	30%	0.767	0.759-0.774
Without the 1st fine-tuning	81%	92%	74%	0.934	0.930-0.939
With all steps included	**91%**	**90%**	**92%**	**0.960**	**0.958-0.962**
Handcrafted features + SVM [42]	79%	73%	83%	0.874	0.869-0.879

Figure 11 shows examples of misclassified cases. Figure 11(a) indicates a discomfort case that is misclassified as a normal case by the automatic system. In this case, there is no significant image feature that was linked to the status. The face itself is also not captured in the frontal view. However, it should be noted that the images are annotated within a video of a time period where the annotator has temporal information about the infant status. Figure 11(b) indicates a comfort case that is misclassified as discomfort status by our automatic system. In this case, the pacifier in front of the face probably confuses our network.

Table 3 summarizes different performances including the classification accuracy and AUCs of training from scratch, training without our strategic first fine-tuning step , and training with all steps included. We also show the result from a traditional approach [42], which is based on handcrafted features combining geometric features and appearance features with a SVM

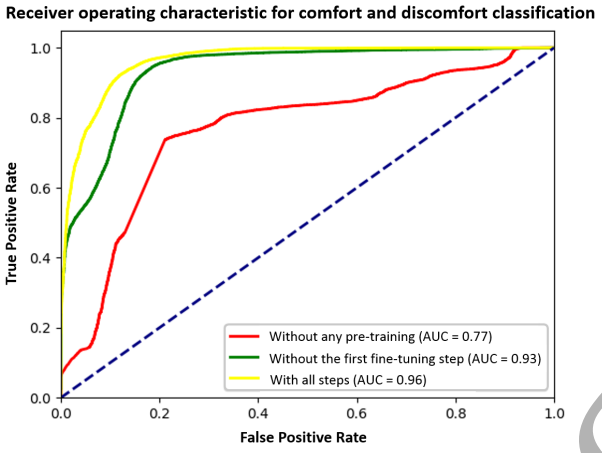


Fig. 12. ROC curves of the proposed method without any pre-training, without the first fine-tuning step, and training with all steps included.

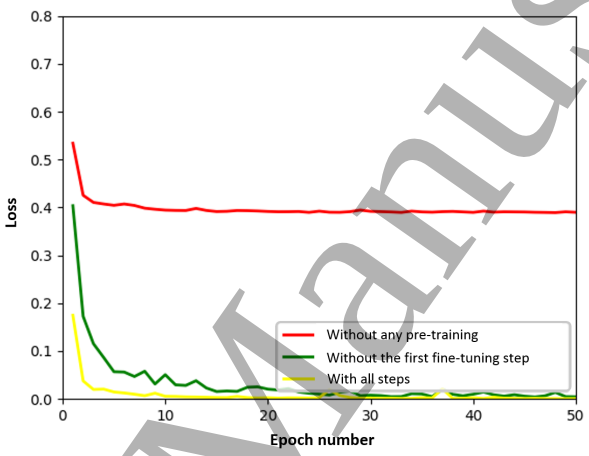


Fig. 13. Loss on the validation data during training epochs without any pre-training, without the first fine-tuning step and training with all steps included.

classifier. For AUCs, the bootstrapping (resampling) approach [5] (1000 bootstrap samples) was used to calculate 95% confidence intervals (CIs).

Figure 12 shows different ROC curves without any pre-training, without the first fine-tuning step and with all the pre-training and fine-tuning steps. The AUC values are 0.77, 0.93 and 0.96, respectively. Figure 13 shows the loss on the validation data of the three training schemes. Our proposed method can quickly reduce the loss during training. Moreover, we can also see that the loss from all three settings are not decreasing after the first 30 epochs which means that 50 epochs are sufficient for training in our case.

The Deep learning CNN are typically referred as a black box for classification while it is difficult to track which features are important for a specific classification task. To understand where deep learning focuses, the Class Activation Map (CAM) [50] can be computed by a weighted sum of the feature maps of the last convolutional layer. CAM can be used to indicate whether our deep learning networks focus on critical facial area or help us understand which region in the face are relevant to this discomfort detection problem.

Table 4. Performance of our proposed method on different imaging factors.

Original test set	zoom factor from 0.8 to 1.2	rotation from -45 degree to 45	contrast from 0.8 to 1.2
AUC	0.96	0.96	0.95
ACC	0.91	0.90	0.91

We observe that quite often the highlighted regions in activation maps are the mouth and eye areas of the face as shown in Figure 14.

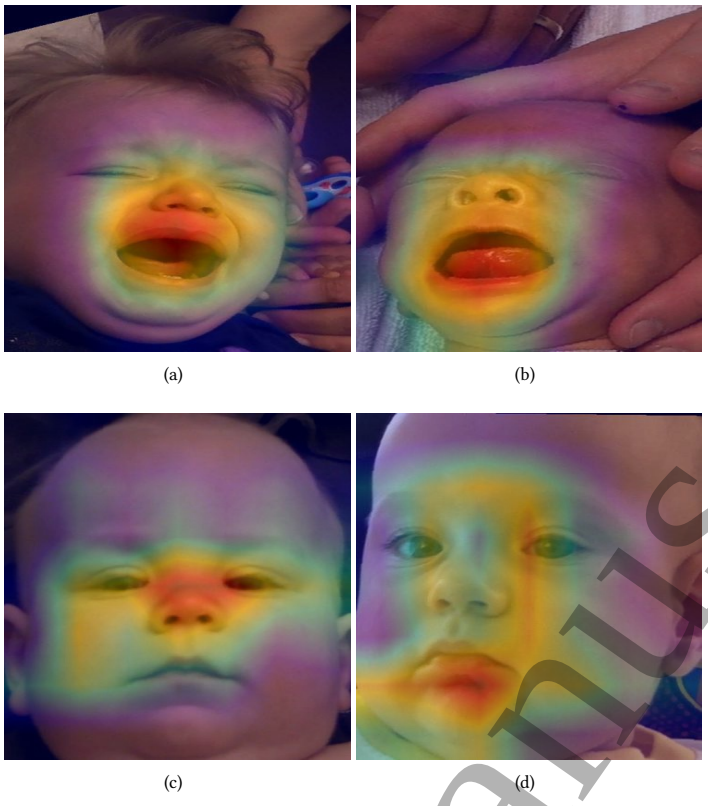


Fig. 14. Examples of discomfort (top row) and comfort (bottom row) faces superimposed by the activation maps. The map highlights the discriminative regions.

We have investigated the performance when randomly zooming in and out from 0.8 up to 1.2, rotating the face from -45 to +45 degrees and changing the contrast from 0.8 to 1.2. The results are shown in Table 4, which demonstrate that our system is robust to different video conditions.

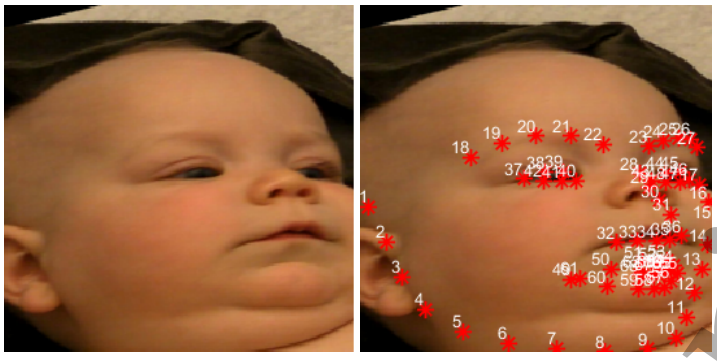
4.5 Video segment classification

We have also included an experiment for segment-based video analysis (5 seconds per segment, i.e. 150 frames) and classification by fusing temporal information. Then, we have computed the mean, maximum and minimum of likelihoods of all frames and obtained AUC values of 0.984 (95% CI: 0.967-0.993), 0.980 (95% CI: 0.958-0.992) and 0.966 (95% CI: 0.939-0.982), respectively.

5 DISCUSSION AND CONCLUSION

We have developed an automatic visual diagnosis system for the classification of discomfort and comfort status of infants in videos. This system applies a deep learning model based on pre-trained DenseNet. Using the strategic fine-tuning steps, our model in combination with two-fold cross-validation, obtained an overall accuracy of 91% on a dataset of 6,534 comfort and 10,303 discomfort video frames from 24 infants. The obtained detection accuracy for comfort and discomfort frames are 90% and 92%, respectively. By fusing individual frame results, the AUC is further improved from 0.96 to 0.98. This indicates that the visual diagnosis system can be potentially used as an alert system to notify the doctors and nurses on the comfort status of the infants. The medical staff can combine the decision of the system together with their own judgments. Furthermore, we have shown that the performance of the deep-learning model is improved when using our proposed strategic fine-tuning steps, involving pre-training with generic people pictures and dataset balancing combined with two-fold cross-validation. Using all refinements, AUC is then substantially increased from 0.77 to 0.96.

The benefits of using an intermediate strategic step of sequential fine-tuning compared to fine-tuning directly on the pre-trained model have been elucidated. Our explanation is that the size of our infant dataset is rather small, which is not suitable for fine-tuning a very large set of parameters of the whole networks at the beginning. Therefore, we first train with a relatively large and similar dataset to tune the networks. In our case, the Shoulder-Pain dataset [25] is quite appropriate,



(a) (b)

(c) (d)

Fig. 15. Examples of false alarms. (a) An failed case from traditional method, (b) with corresponding landmarks at the top. (c) Failed case from the proposed deep learning method and (d) original with its superimposed heatmap.

since it is a labeled facial expression dataset of videos from adults. Although the accuracy in comfort class drops after the first fine-turning step, the overall accuracy and the accuracy in discomfort class increase substantially. In this application, it is more important to have a sensitive system to detect more discomfort frames at the cost of the increase in false positive rate.

The proposed automatic system is very selective. The area under the ROC curve is very high (0.96). From the ROC curve, we can see that we can keep the sensitivity of detecting discomfort status of our vision diagnosis system to be unity, while the specificity is 0.78. It means that our system can identify 78% comfort frames without missing any discomfort frames. From the remaining frames, the medical system can make decisions. In clinical practice, healthcare professionals expect a discomfort detection system that is sensitive to discomfort moments, while producing false alarms as little as possible. The required AUC is not explicit from existing literature by using an automated detection system for this task. Most hospitals use manual assessment by health professionals. However, one possible solution would be that we conduct an observer study in the future to compare the performance of our system with that of experienced medical staff. The AUC and accuracy by our system should compare favorably to medical staff on average.

In our previous work [42], extracting geometric features based on facial landmarks were investigated for discomfort detection. When infants start suffering from discomfort, they tend to squeeze their eyes and stretch their mouths. In order to extract relevant features, the areas of eyes and mouth were calculated by counting the number of pixels inside the polygons surrounded by the landmarks of the eyes and lips. The geometric features achieved an AUC of 0.85 and an accuracy value of 0.78, which is considerably lower than the metrics by using the deep learning-based method.

Both the traditional [42] and the deep learning method made mistakes. Figure 15 shows a typical example where traditional method failed and another example that the deep learning method misclassified comfort as discomfort. The main weakness of traditional features is that it is based on landmark detection. Thus, misplacement of landmarks will affect the accuracy of classification. For deep learning based method, the robustness relies on the availability of sufficient data.

We also compute activation maps for CNNs which allow us to visualize the focused region in a given image, highlighting the discriminative object parts detected by the CNN. In our cases, eyes and mouth are quite often highlighted which confirm regions where handcrafted features are extracted in previous studies.

We implement our deep learning model using the Keras framework. Regarding computation time, the average computation per frame is 0.013 seconds (i.e. 76 fps) using a GTX-980 GPU in the computing system. The execution speed is sufficient for a real-time application.

The features or input from our deep learning system is a static frame, while the temporal features such as movement are not yet used in the algorithm. In future, we will investigate the incorporation of temporal information with the recurrent neural network. For videos, usually, the sound is also recorded. Another direction is to capture and analyze the associated sound information for detection, as it is easy to distinguish an infant is feeling discomfort, when the detection system discovers that the infant is crying. To make deep learning applicable, we extracted 16837 frames from 24 infants. To boost the performance in the future, it is important to recruit more infants.

Acknowledgements. The authors have confirmed that any identifiable participants in this study have given their consent for publication.

REFERENCES

[1] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009.

[2] RE Behrman and A Stith Butler. Institute of medicine committee on understanding premature birth and assuring healthy outcomes board on health sciences outcomes: preterm birth: causes, consequences, and prevention. *Preterm birth: causes, consequences, and prevention, National Academies Press, Washington, DC*, 2007.

[3] Suzanne Brown and Fiona Timmins. An exploration of nurses' knowledge of, and attitudes towards, pain recognition and management in neonates. *Journal of Neonatal nursing*, 11(2):65–71, 2005.

[4] Guiguang Ding, Yuchen Guo, Kai Chen, Chaoqun Chu, Jungong Han, and Qionghai Dai. Decode: deep confidence network for robust image classification. *IEEE Transactions on Image Processing*, 2019.

[5] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[6] Maria Fitzgerald. The development of nociceptive circuits. *Nature Reviews Neuroscience*, 6(7):507–520, 2005.

[7] Maria Fitzgerald, Catherine Millard, and Neil McIntosh. Cutaneous hypersensitivity following peripheral tissue damage in newborn infants and its reversal with topical anaesthesia. *Pain*, 39(1):31–36, 1989.

[8] E. Fotiadou, S. Zinger, W. E. Tjon a Ten, S. Bambang Oetomo, and P. H. N. de With. Video-based facial discomfort analysis for infants. In *Proceedings Volume 9029, Visual Information Processing and Communication*, volume 3, page 9029, 2014.

[9] Ruth E Grunau, Michael F Whitfield, Julianne Petrie-Thomas, Anne R Synnes, Ivan L Cepeda, Adi Keidar, Marilyn Rogers, Margot MacKay, Philippa Hubber-Richard, and Debra Johannesen. Neonatal pain, parenting stress and interaction, in relation to cognitive and motor development at 8 and 18 months in preterm infants. *Pain*, 143(1):138–146, 2009.

[10] Zakia Hammal and Jeffrey F. Cohn. Automatic detection of pain intensity. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 47–52, New York, NY, USA, 2012. ACM.

[11] Jungong Han, L. Hazelhoff, and P.H.N. With, de. *Neonatal monitoring based on facial expression analysis*, pages 303–323. IGI Global, 2012.

[12] Lykele Hazelhoff, Jungong Han, Sidarto Bambang-Oetomo, and Peter H. N. de With. Behavioral state detection of newborns based on facial expression analysis. In Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 698–709, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645, 2016.

[14] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[16] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society, 2014.

[17] Reza Kharghanian, Ali Peiravi, and Farshad Moradi. Pain detection from facial images using unsupervised feature learning approach. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 419–422. IEEE, 2016.

[18] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[19] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[20] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, 16(1):172–187, 2007.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.

[23] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[24] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664–674, 2011.

[25] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain 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] N. Neshov and A. Manolova. Pain detection from facial characteristics using supervised descent method. In *Proc. IEEE 8th Int. Conf. Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 251–256, September 2015.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[28] Joann R Petrini, Todd Dias, Marie C McCormick, Maria L Massolo, Nancy S Green, and Gabriel J Escobar. Increased risk of adverse neurological development for late preterm infants. *The Journal of pediatrics*, 154(2):169–176, 2009.

[29] Kenneth M Prkachin and Patricia E Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008.

[30] Arushi Raghuvanshi and Vivek Choksi. Facial expression recognition with convolutional neural networks. *CS231n Course Projects*, 2016.

[31] Tonse NK Raju, Rosemary D Higgins, Ann R Stark, and Kenneth J Leveno. Optimizing care and outcome for late-preterm (near-term) infants: a summary of the workshop sponsored by the national institute of child health and human development. *Pediatrics*, 118(3):1207–1214, 2006.

[32] R Pillai Riddell and Nicole Racine. Assessing pain in infancy: the caregiver context. *Pain Research and Management*, 14(1):27–32, 2009.

[33] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M Gonfaus, Kamal Nasrollahi, Thomas B Moeslund, and F Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, 2017.

[34] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *International Symposium on Visual Computing*, pages 234–243. Springer, 2013.

[35] Martin Schiavenato, Jacquie F Byers, Paul Scovanner, James M McMahon, Yinglin Xia, Naiji Lu, and Hua He. Neonatal pain facial expression: Evaluating the primal face of pain. *Pain*, 138(2):460–471, 2008.

[36] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[37] Caifeng Shan. An efficient approach to smile detection. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 759 – 764. IEEE, 2011.

[38] Caifeng Shan and Ralph Braspenning. Recognizing facial expressions automatically from video. In *Handbook of ambient intelligence and smart environments*, pages 479–509. Springer, Boston, MA, 20010.

[39] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[40] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang. Automated assessment of children's postoperative pain using computer vision. *Pediatrics*, 136(1):124–131, 2015.

[41] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[42] Yue Sun, Caifeng Shan, Tao Tan, Xi Long, Arash Pourtaherian, Svitlana Zinger, and Peter H. N. With. Video-based discomfort detection for infants. *Machine Vision and Applications*, page 1, August 2018.

[43] Tao Tan, Zhang Li, Haixia Liu, Ping Liu, Wenfang Tang, Hui Li, Yue Sun, Yusheng Yan, Keyu Li, Tao Xu, et al. Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning. *arXiv preprint arXiv:1802.03617*, 2018.

[44] Feng Wang, Xiang Xiang, Chang Liu, Trac D Tran, Austin Reiter, Gregory D Hager, Harry Quon, Jian Cheng, and Alan L Yuille. Regularizing face verification nets for pain intensity regression. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 1087–1091. IEEE, 2017.

[45] R Whit Hall and KJS Anand. Short-and long-term impact of neonatal pain and stress. *NeoReviews*, 6:69–75, 2005.

[46] Xiaojing Xu, Kenneth D Craig, Damaris Diaz, Matthew S Goodwin, Murat Akcakaya, Büşra Tuğçe Susam, Jeannie S Huang, and Virginia R de Sa. Automated pain detection in facial videos of children using human-assisted transfer learning. In *International Workshop on Artificial Intelligence in Health*, pages 162–180. Springer, 2018.

[47] J. Zhao, J. Han, and L. Shao. Unconstrained face recognition using a set-to-set distance measure on deep learned features. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2679–2689, Oct 2018.

[48] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3466–3474, June 2016.

[49] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3474, 2016.

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[51] J. Zhou, X. Hong, F. Su, and G. Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1535–1543, June 2016.

[52] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 84–92, 2016.