Original Article

Camera-based discomfort detection using multi-channel attention 3D-CNN for hospitalized infants

Yue Sun[1], Jingjing Hu[2], Wenjin Wang[1], Min He[2], Peter H. N. de With[1]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands; [2]Department of Electrical Engineering, Hunan University, Changsha, China

Correspondence to: Wenjin Wang. Building Flux P.O. Box 513 5600 MB, Eindhoven, The Netherlands. Email: wwang@tue.nl.

Background: Detecting discomfort in infants is an important topic for their well-being and development. In this paper, we present an automatic and continuous video-based system for monitoring and detecting discomfort in infants.

Methods: The proposed system employs a novel and efficient 3D convolutional neural network (CNN), which achieves an end-to-end solution without the conventional face detection and tracking steps. In the scheme of this study, we thoroughly investigate the video characteristics (e.g., intensity images and motion images) and CNN architectures (e.g., 2D and 3D) for infant discomfort detection. The realized improvements of the 3D-CNN are based on capturing both the motion and the facial expression information of the infants.

Results: The performance of the system is assessed using videos recorded from 24 hospitalized infants by visualizing receiver operating characteristic (ROC) curves and measuring the values of area under the ROC curve (AUC). Additional performance metrics (labeling accuracy) are also calculated. Experimental results show that the proposed system achieves an AUC of 0.99, while the overall labeling accuracy is 0.98.

Conclusions: These results confirms the robustness by using the 3D-CNN for infant discomfort monitoring and capturing both motion and facial expressions simultaneously.

Keywords: Infant discomfort; discomfort detection; video health monitoring; 3D convolutional neural network (3D-CNN)

Submitted Nov 25, 2020. Accepted for publication Mar 29, 2021.
doi: 10.21037/qims-20-1302
View this article at: http://dx.doi.org/10.21037/qims-20-1302

Introduction

Reliable discomfort detection plays a crucial role in appropriate and timely treatment in pediatric clinics. Controlling pain in the newborn period of infants is beneficial in terms of improving physiological, behavioral, and hormonal outcomes (1). Recent findings show that cumulative pain/stress experienced in early life significantly contributes to neurobehavioral development (2). For infants born preterm, neonatal pain-related stress is associated with alterations in both early and later developmental outcomes (3). Significant and long-lasting consequences following pain/stress in the newborn periods can change the central nervous system and responsiveness of the neuroendocrine and immune systems, which can lead to stress at maturity (4,5).

In common practice, there is no universal or standard method to monitor and assess pain. Currently, the stress/comfort levels of hospitalized infants are regularly checked. For pediatric units, multiple validated scoring systems are used. For term neonates, the well-known discomfort/comfort scales are, e.g., the Comfort Scale (6), which considers both observational and physiological factors, and the Neonatal Infant Pain Scale (NIPS) (7), based on interpreting facial expression, crying, breathing patterns, and upper and lower limb movement, etc. For preterm neonates, the Premature Infant Pain Profile (PIPP) score (8)

is a behavioral measure of pain. The scale-based assessment is typically performed by caregivers/clinicians by observing infant behavior for 2–3 minutes. However, the visual assessments are only scheduled a few times a day, which leaves many possibilities of delayed or even missed detections of discomfort status. Another important aspect of detection is that the current procedure is based on the subjective assessment of personnel.

Since a continuous discomfort/pain assessment tool is not available, an automatic monitoring system that can continuously detect discomfort is highly desired to replace the current intermittent manual observation, such as by a camera-based health monitoring system. Emotion recognition is a classic and challenging topic in computer vision. The advanced methods for recognizing emotion or behavior are mostly developed for adults, but a similar principle can be leveraged to measure the comfort/discomfort of infants.

In this paper, we propose a 3D-CNN for capturing temporal and spatial changes in infant facial appearance and body position. The motivation for developing a 3D-CNN method is to exploit both the spatial contextual information (e.g., appearance) and the temporal information (e.g., motion) in a single optimization framework for discomfort/comfort classification. This means that the learned CNN-features incorporate the spatial and temporal features simultaneously. Based on the 3D-CNN architecture, we further investigate different inputs to the network to understand the importance of temporal motion information to discomfort classification, such as 3-channel RGB, 2-channel motion, and 5-channel combination of both. A thorough benchmark is performed between 2D-CNN and 3D-CNN, also between different channel inputs. Eventually, we arrive at an overall best solution that uses five input channels for 3D-CNN with a short time length, namely "multi-channel attention 3D-CNN". The CNN training and validation are performed on a real clinical dataset that is intended for the study of infant discomfort/comfort. The contributions of this paper are summarized as follows. To the best of our knowledge, we are the first to contribute a 5-channel spatial-temporal CNN to learn infant behavior for discomfort detection. Second, we provide a method to integrate motion information with RGB information in order to derive a method for detecting discomfort at higher performance. Third, the proposed 3D-CNN approach has been validated on a clinical dataset. The design of the novel 3D-CNN network for clinical practice is of importance to achieve continuous reliable operation for clinicians.

The remainder of this paper is organized as follows. The following part of Section Introduction provides an overview of related work. Section Methods elaborates the proposed network, and describes the experimental setup. The experimental results are presented in Section Results, and discussed in Section Discussion. Finally, Section Conclusions concludes the paper.

Related work

Studies in emotion recognition have focused on image-based and video-based approaches (9). Video-based approaches have shown improved recognition performance, since they can exploit temporal features and associate those with emotion changes (10). For the applications of 2D-image processing in infant discomfort detection, extensive research has been focused on facial expression recognition. Sun *et al.* (11) proposed a sequential fine-tuning strategy to classify 2D-images from infant videos and achieved an Area Under the Curve (AUC) value of 0.96. By fusing individual frame results, the AUC was further improved from 0.96 to 0.98. Meng *et al.* (12) proposed Identity-Aware Convolutional Neural Network (IACNN) for facial expression recognition. The results showed an accuracy of 71.29% when testing on the CK+ dataset (13). Liu *et al.* (14) upgraded a single CNN to an ensemble of CNNs and the best single subnet achieved 62.44% accuracy while the whole model scored 65.03% accuracy. However, this ensemble approach may be limited in real-time applications. Mollahosseini *et al.* (15) presented a deep neural network architecture for automated facial expression recognition, which consists of two convolutional layers, followed by max-pooling and four Inception layers. The Inception layers increase the depth and width of the network while keeping the computational budget constant. The work was evaluated on the CMU Multi-PIE face database (16) and achieved an accuracy of 94.7%. Uddin *et al.* (17) leveraged a depth-camera-based solution for efficient facial expression recognition, in which for each pixel in a depth image, eight local directional strengths are obtained and ranked. Incremental to different 2D-CNN approaches, Li *et al.* (18) showed the benefits with data augmentation including face cropping and rotation. Zhao *et al.* (19) proposed a novel set-to-set (S2S) distance measure to calculate the similarity between two sets, in order to improve the recognition accuracy for faces with real-

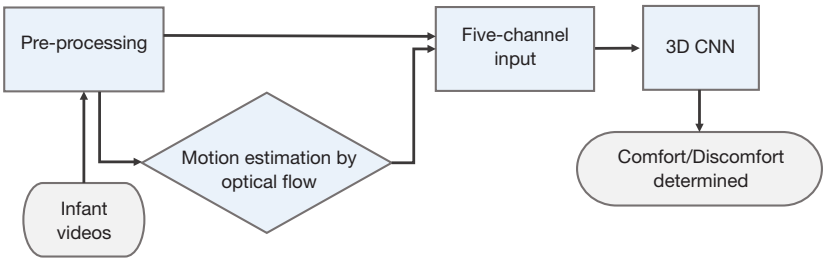


Figure 1 Workflow of our video-based discomfort detection system. 3D-CNN, 3D convolutional neural network.

world challenges. For the S2S distance, the kNN-average pooling is adopted for computing the similarity scores. Luan *et al.* (20) proposed Gabor convolutional networks (GCNs), which utilize Gabor filters as the convolutional filters, such that the robustness of learned features against the orientation and scale changes can be reinforced. Zhang *et al.* (21) developed a new representation learning method, named Structure Transfer Machine (STM), which enables the feature learning process to converge at the representation expectation in a probabilistic way.

In terms of further exploiting temporal information in a video, less attention has been given to facial expression recognition. One recent relevant work is presented by Sun *et al.* (22), where the motion acceleration rate and 18 time- and frequency-domain features were used to characterize motion patterns, leading to an AUC of 0.94 on an infant dataset. Later, the same authors employed optical flow to estimate body motion across video frames to generate feature images, such as Log Mel-spectrogram, Mel Frequency Cepstral Coefficients, and Spectral Subband Centroid Frequency, which were combined by deep CNNs achieving an AUC value of 0.985. On an adult dataset, Zhao *et al.* (23) investigated learning deep facial expression features from the image and optical flow sequences using 3D-CNN, and obtained an average emotion recognition accuracy ranging from 0.56 to 0.76. Jung *et al.* (24) developed a joint network for facial expression recognition, which includes two networks of (I) the deep temporal appearance network (DTAN) and (II) the deep temporal geometry network (DTGN). The CNN-based DTAN is used to extract the temporal appearance feature, while the DTGN is employed for capturing geometric information of facial landmark movements. These two models are further combined to increase recognition performance. The previously discussed work on facial expression and behavioral analysis is still exploratory from nature, and need to be further strengthened.

Methods

In this study, we propose a pipeline (shown in *Figure 1*), which compromises four steps: (I) preprocessing of the input infant videos; (II) optical flow-based motion estimation for estimating body movement between video frames; (III) combination of the images from three RGB channels with two motion channels derived from Step (II). Thus, in total five channels are used as input to the classification network; (IV) application of a 3D-CNN for the binary classification of comfort and discomfort, which embeds a motion-attention module (two motion channels). The 3D-CNN extracts discriminative spatio-temporal representations for different infant status, and finally, a decision of comfort/discomfort is assigned to each video segment. The details of each step are elaborated in the following subsections.

Preprocessing

All video frames are first cropped to remove superfluous information in the image margins. The original image size of the recorded frame is 720×1,280 pixels. The size is decreased to 501×751 pixels after removing pixels along each margin. The image is further down-sampled to 100×150 by nearest-neighbor interpolation (25) to fit the required dimension for the input of the network.

For each down-sampled video frame, we apply Gaussian weighting to suppress the background content that is irrelevant for analysis, for example, background pixel changes caused by caregivers or parents moving around infants for care-handling. In our recording scenario, infants were always located at the central area of the video frames. The Gaussian mask is thereby applied as a weighting function to highlight the central area and suppress boundaries (background area). More specifically, the 2D Gaussian mean locations are directed to the center of the image, while the Gaussian standard deviations on the

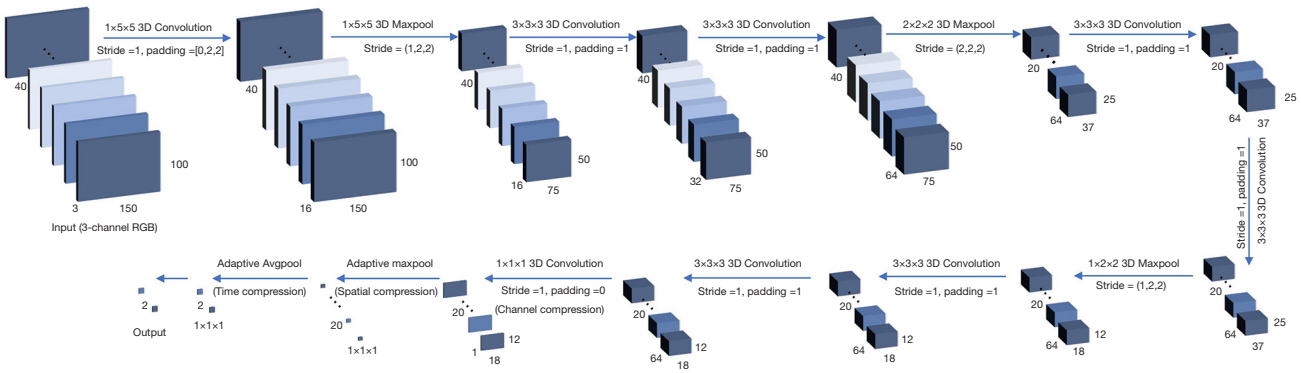


Figure 2 Architecture of the proposed 3D-CNN, which is developed as one deep processing chain. 3D-CNN, 3D convolutional neural network.

horizontal and vertical directions are empirically set to 40 and 80 based on the frame dimension.

The original frame rate of the recorded videos is 30 frames per second (fps). To reduce the computation load and memory cost, each video is sub-sampled to half of the original frame rate (i.e., 15 fps) by keeping the odd-indexed frames and skipping the even-indexed frames. The lower sampling rate also increases the pixel movement between re-sampled adjacent frames and enables better motion analysis.

For each video sequence, we use a temporal sliding window with a fixed length of *M*-frames (i.e., containing frames for *M*/15 seconds) and a sliding stride of *N* frames. Thus, for each window, *M* frames are used as input clip for the classification network.

3D-CNN model

CNNs enable the automatic and deep learning of feature expressions directly from the input data (e.g., images and videos). However, a number of existing CNNs are only capable of handling 2D inputs due to the inherent network structure. The 2D-CNNs are limited to the spatial information, while the temporal information across the video frames is not exploited. The recently proposed 3D-CNNs extract features from both the spatial and temporal domains by performing the 3D convolution and 3D pooling. This approach is able to capture both the object appearance/contextual information and motion information in a single optimization such that the generated features resemble the spatial and temporal semantics (26).

3D-CNNs are extension of 2D-CNNs. As compared to 2D-CNNs, challenges for 3D-CNNs are the larger memory footprint and higher dimensionality. These exacerbate the intensive computation cost for the network inference. To this end, we propose to use a 3D-CNN model that not only considers the spatio-temporal information but also computational efficiency. We use the backbone of the 3D-CNN model described in (27). The input of our network is an image sequence, followed by several convolution and pooling layers as shown in *Figure 2*. In the figure, we use the RGB image as the input to illustrate the 3D-CNN architecture, but this is later extended to a 5-channel approach for feature extraction.

The first convolutional layer contains a number of kernels with the size of 1×5×5, which only convolves the input data for the spatial information in a single frame. The following six convolutional layers with the kernel size of 3×3×3 convolve both temporal and spatial dimensions. It has been shown that a deep net with small filters like 3×3 outperforms a shallow net with larger filters (28), i.e., small receptive fields of 3×3 convolution kernels with deeper architectures yield better results. The previous consideration motivates our choice for the small convolutional filter 3×3×3.

The last convolutional layer is used for channel compression, and then the space is compressed by max-pooling. Max-pooling is generally favored for classification tasks, since it leads to faster convergence and better generalization, while it also retains the most significant and translational invariant information from the convolutional layers. Finally, average-pooling is used for compressing temporal information because the weight of each image in the time series is the same. It finally leads to one confidence

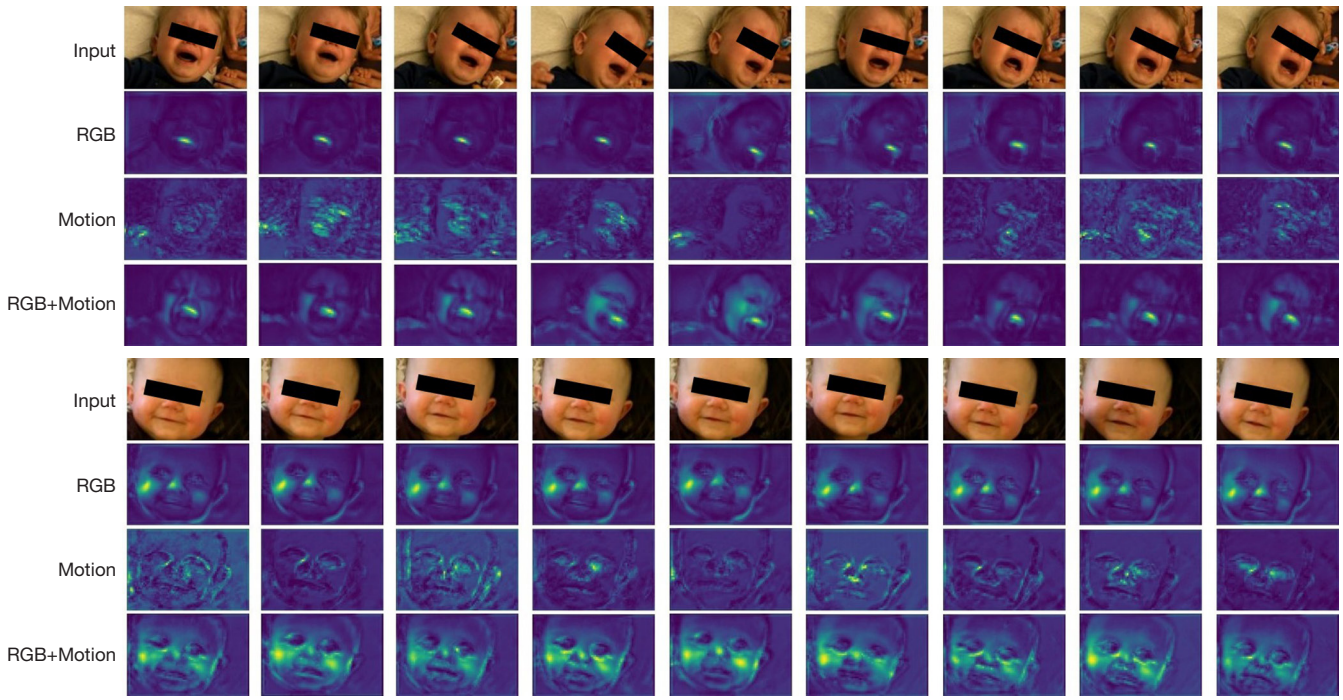


Figure 3 Visualization of our 3D-CNN model. Examples of feature maps for (A) discomfortable and (B) comfortable cases. The first row is the original sequence diagram, and the second shows the feature maps when training only on RGB. The third row is for only using optical flow as input for learning, and the bottom row shows the case when combining the five channels (RGB and optical flow). 3D-CNN, 3D convolutional neural network.

value that determines the label of comfort or discomfort.

After the second convolution layer, the size of each frame is halved, and the number of channels is increased from 3 (RGB) to 64 (number of kernels). For each input video segment, the total number of frames is *M* and each frame includes three channels (RGB). For each frame, the pixel value in feature maps is the sum of squares from all the channels at the corresponding pixel location. Therefore, each video segment generates a group of feature maps.

The obtained outputs of the feature maps from the second convolution layer reveal the motion information, which is from multiple adjacent frames in the original input. *Figure 3* exemplifies the feature maps, which demonstrate that the network can highlight the meaningful motion area of the infants, especially the eye/mouth/nasolabial furrow regions for characterizing their facial expression.

Finally, we specifically disclose the configurations of network functions and optimization. The Adam optimizer is employed for learning. Binary Cross-Entropy (BCE) with Logits Loss (BCEWLL) is used as the loss function. BCE is a cross-entropy suitable for binary classification, which is a special case of multi-class classification softmax cross entropy. For CNN training, we empirically set the number of epochs (*m*) to 800.

Multi-channel attention model

Motion estimation using optical flow

To estimate the motion of infants, we employ optical flow for calculating the motion vectors at the down-sampled pixel level. Pixel-based motion vectors are calculated for each video frame between two consecutive frames, using the dense optical flow of Farnebäck *et al.* (29). It uses quadratic polynomials to estimate the motion between two consecutive frames. Polynomial expansion is employed to approximate pixel intensities in the neighborhoods of the frames. A pyramid decomposition is used to handle large pixel motions, including distances larger than the neighborhood size. The tracking of motion begins at the lowest resolution level and continues until convergence.

Between two consecutive video frames, the optical flow provides motion derivatives. In our study, we compute two motion matrices, which are the velocity magnitude along the horizontal and vertical directions, as illustrated in the

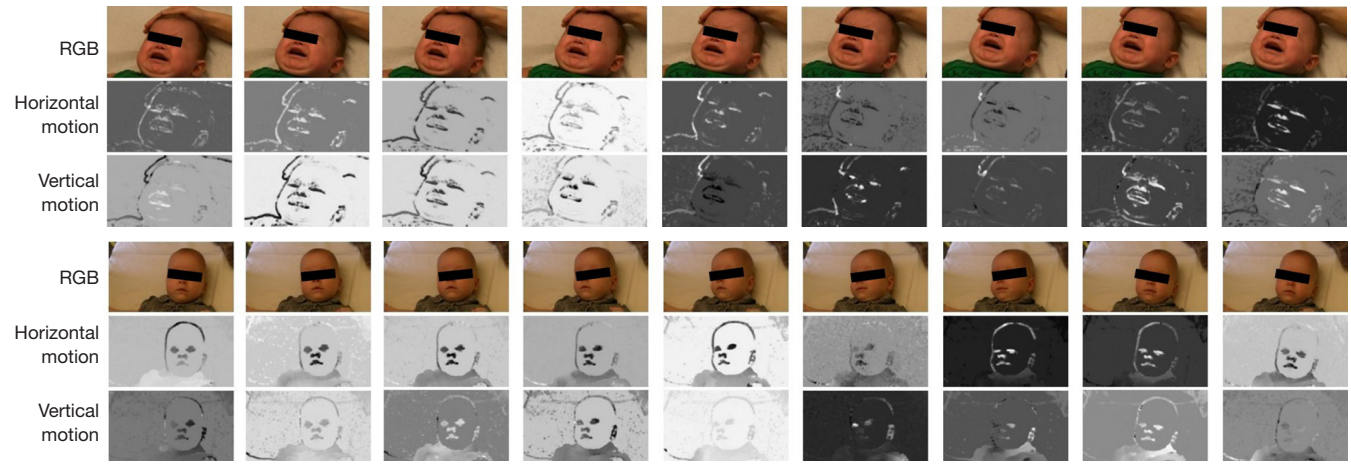


Figure 4 Examples of optical flow images obtained by Farnebäck's method for two infants with a sampling interval of 0.2 s shown in (A) and (B). For each baby, the top row shows the original RGB frames, and the second and third rows are optical flow matrices for horizontal and vertical directions.

second and third rows of *Figure 4A,B*, respectively.

5-channel attention model

We propose an attention-based model to guide the 3D-CNN networks to focus on the image areas where motion may occur. The dense optical flow highlights the motion areas of infant bodies. After calculating the optical flow, we add the two motion matrices containing the motion magnitudes of horizontal and vertical directions as additional channels to the network input.

Figure 3 shows the feature maps generated by different training inputs, which are 3-channel RGB, 2-channel optical flow, and 5-channel combined input. It is clear that the optical flow channels highlight the movement from the videos. By leveraging the information from the optical flow channels, the attention from the neural networks focuses on the facial area in the images. We now further investigate the performance of the fused 5-channel mechanism in the benchmark.

Experimental setup

Clinical dataset

The study was conducted with videos recorded at the Màxima Medical Center (MMC) in Veldhoven, The Netherlands, by a handheld high-definition camera (Xacti VPC-FH1BK). The study was approved by the ethics board of MMC. For each infant in the database, written consent was obtained from the parents. Videos of 24 infants were recorded. The infants' faces were recorded when experiencing various stress/pain moments, including clinical treatments of heel prick, placing an intravenous (IV) line, venipuncture, vaccination, post-operative pain, and the discomfort moments caused by a diaper change, feeling hungry or crying for attention. For 10 out of 24 infants, the relaxed comfort moments of resting or sleeping were also recorded. In conclusion, for these 10 infants both comfort and discomfort status were presented. Four infants only have the comfort moments recorded. For the remaining 10 infants, only discomfort moments were recorded. Therefore, the video segments contain 1 to 2 types of emotion status per subject. The duration of each video varies from less than 1 minute to several minutes. The age of the 24 recorded infants ranges between 2 days and 13 months old. Three infants born prematurely, under 37 weeks at the time of recording.

The resolution of the recorded video frames is 1,920×1,080 pixels, and the original frame rate is 30 fps. The videos were recorded under uncontrolled lighting conditions, i.e., the general office lighting conditions. The label of comfort/discomfort for each frame is manually annotated, based on the consensus of two clinical experts. A total of 55 video segments from 24 infants were selected, where 19 segments present comfort, and the remaining 36 are discomfort cases. There are more discomfort samples in our dataset than comfort samples. The reason is that the data collection performed by the clinicians in the hospital has been focused on discomfort moments of the infants.

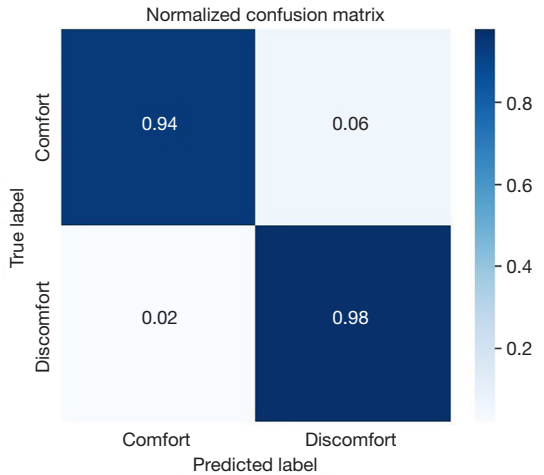


Figure 5 Normalized confusion matrix when directly training on 3-channel RGB images.

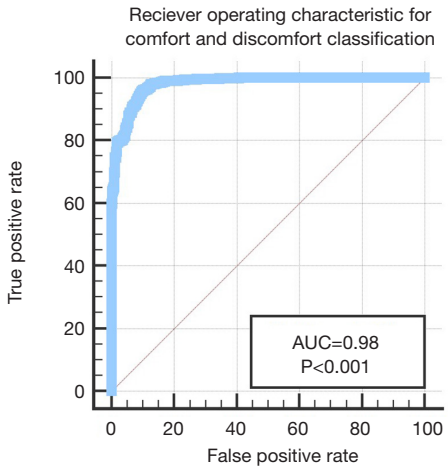


Figure 6 ROC of the proposed method when directly training on 3-channel RGB images. ROC, receiver operating characteristic.

Evaluation metrics

The classification was evaluated by measuring the classification accuracy and confusion matrices. The receiver operating characteristic (ROC) curves are also used for visualizing the performance with the corresponding area under the ROC curve (AUC) values.

Learning protocol

Learning schemes

We exploited different training schemes to investigate the networks, which are:

- 3-channel RGB: training is directly performed on the 3- channel RGB images of our infant dataset, as described in Section Methods.
- 2-channel motion: training is based on the 2-channel motion-derivative images estimated by dense optical flow.
- 5-channel RGB and motion: the hybrid 5-channel input including RGB and motion planes as described in Section Methods.

Window-size tuning

We also investigated the influence of the temporal sliding-window length on the 3D-CNN model, in terms of classification AUC, accuracy, and execution time. The window length was tuned from 1 frame (i.e., the 2D case) to 15 frames (2 s), 30 frames (3 s), …, 90 frames (6 s). The experiments were carried out on a GeForce RTX-2080 GPU and a Xeon E5-2680 V2 CPU.

Results

Evaluation metrics

Learning scheme 3-channel RGB

Figure 5 shows the normalized confusion matrix when the training is directly performed on 3-channel RGB images. The obtained overall labeling accuracy is 0.96. The detection accuracy of comfort and discomfort is 0.94 and 0.98, respectively. *Figure 6* represents the corresponding ROC with the AUC value of 0.98.

Learning scheme 2-channel motion

The measured AUC is 0.73 when training the model only on the 2-channel optical flow images.

Learning scheme 5-channel RGB and motion

The normalized confusion matrix of fine-tuning using 5 channels is shown in *Figure 7*. The model achieved the highest AUC of 0.99 when training on the input of five channels (see *Figure 8*). The overall accuracy delivers robust and consistent information on reaching the best value of 0.98.

Table 1 shows the comparison of the results achieved by the proposed method with existing infant discomfort detection methods of (I) conventional handcrafted features combined with a support vector machine (SVM) (30), and (II) image-level deep learning-based comfort/discomfort classification using fine-tuning strategies (11).

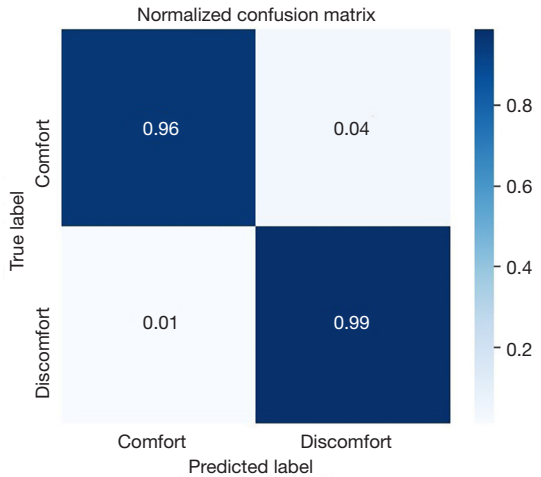


Figure 7 Normalized confusion matrix for 5-channel training.

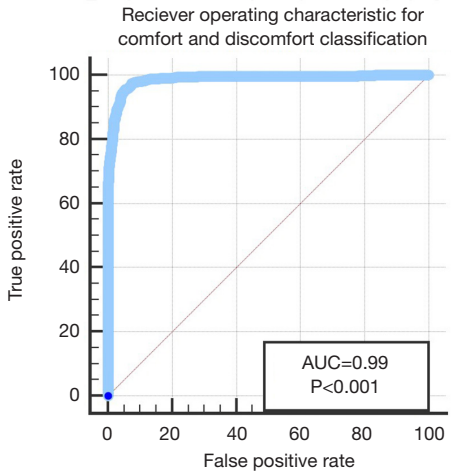


Figure 8 ROC curve of 5-channel training. ROC, receiver operating characteristic.

Table 1 Comparison with existing infant monitoring methods

Method	Overall accuracy	AUC
Handcrafted features + SVM (30)	0.85	0.87
Image-level deep learning with fine-tunings (11)	0.91	0.96
Proposed multi-channel 3D-CNN method	0.98	0.99

AUC, area under the ROC curve; SVM, support vector machine; 3D-CNN, 3D convolutional neural network.

Window-size tuning

The results mentioned above are all based on the window length of 15 frames (2 s) and a step size (stride) of 5 frames (around 167 ms). We further explore the effects by changing the window size. *Figure 9* presents the AUCs for three learning schemes with different window lengths. *Figure 10* shows the accuracy values during the window-size tuning procedure.

Execution time

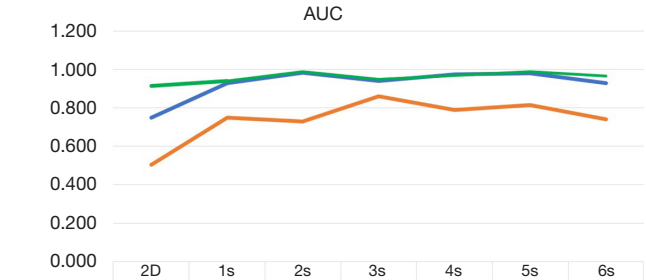
Based on different lengths of the video clips, the corresponding execution times for testing each video clip are summarized in *Table 2*.

Discussion

We provided a neat end-to-end video processing solution without requiring the conventional front-end steps of face detection and tracking. In general practice, the partial/full face occlusion is likely to happen, especially during the infant physical care. Our method is more suitable to handle these challenging moments as compared to traditional face detection/tracking-based methods.

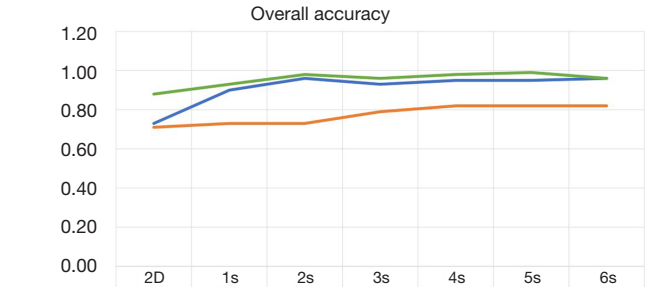
In real clinical practice, healthcare professionals desire to have a discomfort detection system that is sensitive to discomfort moments, while suppressing false alarms as much as possible. From existing literature, it can be derived that the required AUC is not well defined for this task. There is no contact sensor to measure and quantify the discomfort of infants. It is mainly based on the observation, experience, and judgement of clinicians, which unfortunately lacks a medical gold standard. In the future, we can conduct an observer study to compare the performance of our system with that of experienced medical staff.

The best performance is achieved when the sliding window length is set to 2 seconds and the step size is set to one-third of a second. This means that (I) the latency between the start of the system and the first measurement can be as short as 2 seconds, and (II) the monitored infant status can be refreshed every one-third of a second plus the



	2D	1s	2s	3s	4s	5s	6s
RGB	0.748	0.927	0.981	0.939	0.975	0.980	0.929
Optical flow	0.505	0.750	0.730	0.859	0.790	0.813	0.741
5 channels	0.913	0.939	0.988	0.947	0.968	0.987	0.965

Figure 9 AUCs for using different clip lengths and training schemes. AUC, area under the ROC curve.



	2D	1s	2s	3s	4s	5s	6s
RGB	0.73	0.90	0.96	0.93	0.95	0.95	0.96
Optical flow	0.71	0.73	0.73	0.79	0.82	0.82	0.82
5 channels	0.88	0.93	0.98	0.96	0.98	0.99	0.96

Figure 10 Accuracy for using different clip lengths and training schemes.

Table 2 Execution times of the three training methods for using different lengths of video clips

Training method	2D	1 s (s)	2 s (s)	3 s (s)	4 s (s)	5 s (s)	6 s (s)
3-channel RGB	0.064	0.074	0.086	0.097	0.108	0.111	0.128
2-channel motion	0.035	0.073	0.081	0.091	0.099	0.104	0.111
5-channel RGB + motion	0.064	0.078	0.097	0.109	0.127	0.139	0.159

execution time for making a decision on the present clip. The testing execution speed shown in *Table 2* indicates the required processing time for an unseen video clip, which is about 0.1 seconds. This indicates that the infant discomfort detection can be implemented as a real-time clinical application. Compared to 2- or 3-channel schemes, the 5-channel scheme sacrifices some computing time, but the added extra time is negligible.

When comparing the performance of different training schemes, the AUC and overall accuracy are the highest when applying the learning scheme of 5-channel RGB and motion, which confirms the effectiveness of the 5-channel attention-based network. The two channels of optical flow images provide the regions/boundaries with strong motion information, which serve as a beacon that gears the network to focus on the movement. The motion patterns of infants are clinically important for assessing comfort or discomfort. From the feature maps in *Figure 3*, we confirm the conclusion that by incorporating the information from the optical flow channels, the attention from the neural networks is dominantly focused on the facial area of the infant.

When the sliding-window length is set to one frame, this training/testing procedure becomes 2D processing. Obviously, its performance is significantly reduced, as compared to the 3D processing. For all three training schemes, the AUC and accuracy values improve along with the increasing window size, and appear to be saturated after 2 seconds. This may be caused by the limited valid information from the samples or constraints by the available training sample size, which shall be further investigated by collecting more infant data in the future. The reason for the saturated performance could also be that the discomfort information (facial expression or body motion) is mostly spontaneous/abrupt changes that typically happen in a very short time interval, i.e., high-frequency temporal information to be captured in a short time window. A long time window may include and highlight low-frequency information (or slow changes like motion drift or hand-held camera motion) that are less relevant for comfort/discomfort classification.

Regarding the complexity of the proposed 3D CNN, the total number of the parameters is 514,657, yielding a compact computational model. The execution time for making a decision on an unseen clip is 0.097 seconds using the sliding window of 2 seconds, which results in the best AUC of 0.99. As a conclusion, the complexity of the 3D CNN is lower than expected, while it gives a very high performance.

In the future, our model can be improved to identify

discomfort grades by changing the two-class output to multiple classes, or even a regression layer to predict the significance of discomfort. Our future work may also include extending the 5-channel input to higher dimensions by fusing information from different sensor modalities such as the depth information from a 3D sensor (e.g., time-of-flight camera). In addition to the fusion of color intensity signals and motion signals, we also consider fusing the contextual information and physiological information for joint classification, since physiological variables (e.g., heart-rate, heart rate variability, and respiration rate) are also possible to be measured from the videos.

Conclusions

This paper proposes a video monitoring system that provides continuous and contactless assessment of discomfort for infants. The system is validated by real clinical infant data with expert annotations. In this study, we have investigated the benefit of using optical flow measurement to draw the attention of 3D-CNNs. The system aims to alert caregivers/clinicians immediately when infants start suffering from discomfort. The proposed system can monitor infant status continuously by classifying the video frames into either comfort or discomfort, which fills the gaps of the current intermittent manual observation. Moreover, the proposed method also has the potential to be implemented as an infant-care tool for family use on a longer term. The system alarm is triggered by the detection of discomfort status, which will notify clinical staff for timely and appropriate treatment. Thus, the system serves to prevent fatal events and eventually improves the early development of infants.

Acknowledgments

Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/qims-20-1302). The authors have no conflicts of interest to declare.

Ethical Statement: The study was approved by the ethics board of Màxima Medical Center (MMC). For each infant in the database, written consent was obtained from the parents.

References

1. Witt N, Coynor S, Edwards C, Bradshaw H. A guide to pain assessment and management in the neonate. Curr Emerg Hosp Med Rep 2016;4:1-10.
2. Cong X, Wu J, Vittner D, Xu W, Hussain N, Galvin S. The impact of cumulative pain/stress on neurobehavioral development of preterm infants in the NICU. Early Hum Dev 2017;108:9-16.
3. Valeri BO, Holsti L, Linhares MB. Neonatal pain and developmental outcomes in children born preterm: a systematic review. Clin J Pain 2015;31:355-62.
4. Grunau RE, Tu MT. Long-term consequences of pain in human neonates. In: Anand KJ, Stevens B, Mcgrath P. editors. Pain in neonates and infants. 3rd ed: London, Great Britain: Elsevier; 2007:45-55.
5. Page GG. Are there long-term consequences of pain in newborn or very young infants? J Perinat Educ 2004;13:10-7.
6. Ambuel B, Hamlett KW, Marx CM, Blumer JL. Assessing distress in pediatric intensive care environments: the COMFORT scale. J Pediatr Psychol 1992;17:95-109.
7. Lawrence J, Alcock D, McGrath P, Kay J, MacMurray SB, Dulberg C. The development of a tool to assess neonatal pain. Neonatal network: NN 1993;12:59-66.
8. Stevens B, Johnston C, Petryshen P, Taddio A. Premature infant pain profile: development and initial validation. Clin J Pain 1996;12:13-22.
9. Dhall A, Ramana Murthy O, Goecke R, Joshi J, Gedeon T. Video and image-based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction; 2015:423-6.
10. Fan Y, Lu X, Li D, Liu Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International

Conference on Multimodal Interaction; 2016:445-50.
11. Sun Y, Shan C, Tan T, Tong T, Wang W, Pourtaherian A, de With PHN. Detecting discomfort in infants through facial expressions. Physiol Meas 2019;40:115006.
12. Meng Z, Liu P, Cai J, Han S, Tong Y. Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE; 2017:558-65.
13. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE; 2010:94-101.
14. Liu K, Zhang M, Pan Z. Facial expression recognition with CNN ensemble. In: 2016 international conference on cyberworlds (CW). IEEE; 2016:163-6.
15. Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV). IEEE; 2016:1-10.
16. Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multi-PIE. Proc Int Conf Autom Face Gesture Recognit 2010;28:807-13.
17. Uddin MZ, Khaksar W, Torresen J. Facial expression recognition using salient features and convolutional neural network. IEEE Access 2017;5:26146-61.
18. Li K, Jin Y, Akram MW, Han R, Chen J. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Visual Comput 2020;36:391-404.
19. Zhao J, Han J, Shao L. Unconstrained face recognition using a set-to-set distance measure on deep learned features. IEEE Trans Circuits Syst Video Technol 2017;28:2679-89.
20. Luan S, Chen C, Zhang B, Han J, Liu J. Gabor convolutional networks. IEEE Trans Image Process 2018;27:4357-66.
21. Zhang B, Yang W, Wang Z, Zhuo L, Han J, Zhen X. The structure transfer machine theory and applications. IEEE Trans Image Process 2019;29:2889-902.
22. Sun Y, Kommers D, Wang W, Joshi R, Shan C, Tan T, Aarts R, Carola VP, Peter A, Peter HNDW. 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 eMedicine and Biology Society (EMBC). IEEE; 2019:5995-9.
23. Zhao J, Mao X, Zhang J. Learning deep facial expression features from image and optical flow sequences using 3D CNN. Visual Comput 2018;34:1461-75.
24. Jung H, Lee S, Yim J, Park S, Kim J. Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision; 2015:2983-91.
25. Parker J, Kenyon RV, Troxel DE. Comparison of interpolating methods for image resampling. IEEE Trans Med Imaging 1983;2:31-9.
26. Ji S, Xu W, Yang M, Yu K. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Trans Pattern Anal Mach Intell 2013;35:221-31.
27. Yu Z, Li X, Zhao G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: Proc. BMVC; 2019:1-12.
28. Simonyan K, Zisserman A. Very deep convolutional networks for large- scale image recognition. CoRR, abs/1409.1556, 2014; 1, 2, 3, 5.
29. Farnebäck G. Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image Analysis. Springer; 2003:363-70.
30. Sun Y, Shan C, Tan T, Long X, Pourtaherian A, Zinger S, de With PHN. Video-based discomfort detection for infants. Mach Vis Appl 2019;30:933-44.

Cite this article as: Sun Y, Hu J, Wang W, He M, de With PHN. Camera-based discomfort detection using multi-channel attention 3D-CNN for hospitalized infants. Quant Imaging Med Surg 2021;11(7):3059-3069. doi: 10.21037/qims-20-1302

