Automated Infant Monitoring based on R-CNN and HMM

Cheng Li[1][a], A. Pourtaherian[1][b], L. van Onzenoort[2] and P. H. N. de With[1][c]
[1]*Eindhoven University of Technology, Eindhoven, The Netherlands*
[2]*Maxima Medical Center, Veldhoven, The Netherlands*

Keywords: Automated Infant Monitoring, R-CNN, HMM, GERD Diagnosis.

Abstract: Manual monitoring of young infants suffering from reflux is a significant effort, since infants can hardly articulate their feelings. This work proposes a near real-time video-based infant monitoring system for the analysis of infant expressions. The discomfort moments can be correlated with a reflux measurement for gastroesophageal reflux disease diagnose. The system consists of two components: expression classification and expression state stabilization. The expression classification is realized by Faster R-CNN and the state stabilization is implemented with a Hidden Markov Model. The experimental results show a mean average precision of 82.3% and 83.4% for 7 different expression classifications, and up to 90% for discomfort detection, evaluated with both clinical and daily datasets. Moreover, when adopting temporal analysis, the false expression changes between frames can be reduced up to 65%, which significantly enhances the consistency of the system output.

1 INTRODUCTION

Young infant expression analysis is a difficult and important task within the field of pediatrics, since the verbal ability of young infants is limited or not yet even developed. Over the years, pain assessment tools of infants have been developed based on subjective descriptors such as facial expressions. However, these systems require practitioners to observe infants over time, which is laborious and time-consuming. Likewise, continuous monitoring and nursing of infants by professionals would also be too expensive. To address this, an automated video-based infant monitoring system could analyze expressions as an auxiliary assessment tool. This approach would differentiate from other methods, such as heart-rate monitoring, since video-based expression analysis has a non-interventional character. Besides, the discomfort moments detected by such a system can be correlated with disease diagnosis. For example, gastroesophageal reflux disease (GERD) is a common disease for young infants and is visible by unexpected vomiting of the infant, which causes discomfort and sometimes pain. Such an event can be measured and combined with the generic monitoring for diagnostic purposes, which is attractive by pediatricians. However, to design such a system, several challenges have to be resolved.

First, for automated infant observation, the system should perform face detection, even when the infant has large deviating head poses far away from the camera view. Second, facial expressions of infants are significantly different from those of adults. Hence, most of the state-of-the-art face detection and expression analysis methods are trained with adult datasets and therefore typically fail in pediatric applications. As a result, a specific approach is required for infants. Third, public infant datasets are rare in contents, so that training a deep learning-based classifier with little data will not yield reliable results and may become over-fitted. Fourth, the trained classifier should be able to detect expressions when faces are partially occluded by objects. Finally, a steady expression-state output of a system is also important, since a false alarm can bring misleading information to caregivers and doctors for disease diagnosis. Moreover, frequent false chirps of an infant monitoring system are also bothering, and therefore make the system perceived as less reliable, even when the discomfort detection accuracy is overall high. To overcome the aforementioned challenges, we propose to use CNN networks because they have proven to be successful in multiple detection problems offering a high accuracy and they

[a] https://orcid.org/0000-0003-2900-637X
[b] https://orcid.org/0000-0003-4542-1354
[c] https://orcid.org/0000-0002-7639-7716

Li, C., Pourtaherian, A., van Onzenoort, L. and N. de With, P.
Automated Infant Monitoring based on R-CNN and HMM.
DOI: 10.5220/0010299605530560
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 553-560
ISBN: 978-989-758-488-6

can be robust when properly trained (Mueller et al., 2017; Long et al., 2015). From these CNNs, we consider that an automated infant monitoring system can be constructed using Faster R-CNN (Ren et al., 2017), which can handle the variation from the above challenging situations. Faster R-CNN was adopted because it requires less memory than LSTM (Donahue et al., 2017) and the processing is less complex. Additionally, instead of expensive LSTM, we employ a dynamic Hidden Markov Model (HMM) in combination with Faster R-CNN which models changes of the expression over time, so that the frame-based dynamics of the expression classification become more stable and avoid drastic changes. This combination enables near real-time operations and provides both reliable detection and temporally stable behavior, which is novel extension to this application.

In this work, we make the following contributions to automated infant observations.

1. *Multi-facial Expression Classification:* Infant expressions are subtle and instinctive, which explains why we distinguish infant expressions into seven states, instead of a binary detection.
2. *Hybrid CNN Model:* We propose to combine Faster R-CNN with HMM to model the dynamic changes of the infant expression between video frames over time, which improves the robustness of expression classification and reduces time-domain complexity considerably.
3. *Large-scale Validation:* The discomfort detection system is evaluated with large-scale datasets, which encompass the datasets collected from both clinical environments and normal daily life.
4. *Consistent Expression Status:* To our best knowledge, the temporal status consistency of an infant monitoring system output is for the first time evaluated in this work. The experimental results show a significant improvement of the discomfort detection accuracy as well as the output consistency when using Faster R-CNN and HMM.

The remainder of this paper is organized as follows. Section 2 briefly introduces some related work. Afterwards, the design details of the system are explained in Section 3. The experimental results for infant expression analysis are provided in Section 4. Finally, Section 5 presents conclusions.

2 RELATED WORK

The absence of regular pain assessment for infants makes pain for infants normally under treated. Unfortunately, no gold standard or universal approach for infant pain assessment is available in pediatric fields. For decades, researchers have paid significant attention to devising multiple validated pain scoring systems for facilitating the objective measurement of pain. A thorough introduction of pain assessment methodologies for neonates is provided in (Witt et al., 2016). These assessment tools use facial expressions as a main indicator, such as the revised FACES pain scale (Hicks et al., 2001), FLACC (Jaskowski, 1998), the Wong-Baker Faces scale, neonatal facial coding system (NFCS) (Witt et al., 2016) and the 10-cm visual analog scale, and are widely used in healthcare settings to assess patient pain.

Inspired by these pain assessments, some automated pain detection systems for neonates have been studied based on facial expression analysis. In (Fotiadou et al., 2014), a semi-automated system is proposed for discomfort detection, which adopts an Active Appearance Model (AAM) for facial appearance modeling. Zhi *et al.* (Zhi et al., 2018) proposed an automatic pain detection for infants by using geometric features. After that, a Support Vector Machine (SVM) is utilized to distinguish pain from no pain. In (Li et al., 2016), it is proposed to exploit an automated classification model, based on appearance features extracted by local binary patterns. Sun *et al.* (Sun et al., 2018) presented a discomfort detection system based on a template matching method, in which neutral faces of the specific subject are used as a template. Frames containing different expressions compared with the template are classified as discomfort. However, all these mentioned methods can only handle situations where infants show their frontal faces without occlusions, hence giving limited robustness.

Recently, Convolutional Neural Networks (CNNs) have become prevalent because of their increased performance. Tavakolian *et al.* (Tavakolian and Hadid, 2018) proposed a pain expression intensity estimation based on CNNs and a binary coding. Lin *et al.* (Lin et al., 2018) have proposed a CNN-based expression classifier trained with data augmentation for adults, which is similar to our work. However, these works targeted on adults are not applicable to infants as already mentioned. Therefore, we propose an infant expression analysis algorithm using Faster R-CNN and a temporal model for realizing temporal stability in decision making. Although CNNs can also applied in the temporal domain, this approach is hampered by the high intrinsic complexity. Therefore we explore an alternative, since our aim is to implement a real-time automated infant monitoring system suitable for clinical practice.

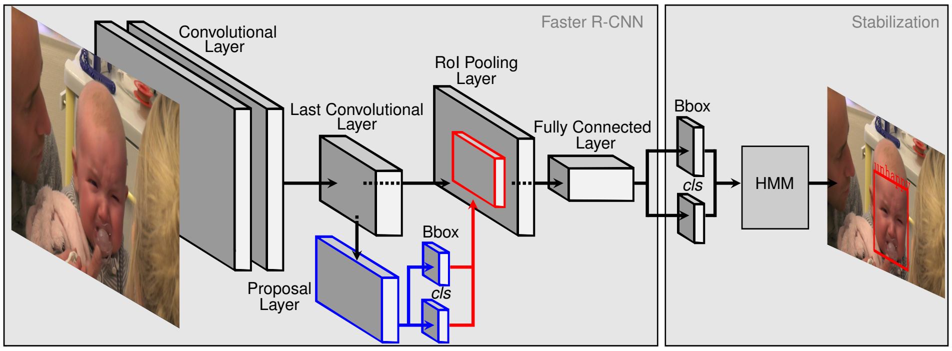


Figure 1: Flowchart of the system design for infant discomfort analysis based on Faster R-CNN and HMM.

3 INFANT DISCOMFORT DETECTION

3.1 R-CNN for Expression Detection

For expression detection and classification, the Faster R-CNN (Ren et al., 2017) framework combined with a VGG-net architecture (Simonyan and Zisserman, 2014) is adopted, since Faster R-CNN yields state-of-art performance for object detection. The training input of the expression detection network is represented as a tuple (I_m, g_i, cls), where I_m is the full-image frame containing the infant face and background, g_i denotes the ground truth of a bounding box encompassing the infant face and cls indicates the ground-truth label for expressions of that bounding box. In this application, cls corresponds to one of the following expression states: Discomfort/pain, Unhappy, Neutral, Sleep, Joy, Open mouth and Pacifier, defined as in (Sullivan and Lewis, 2003). For training, the Regions of Interests (RoIs) that have an intersection over union (IoU) with the ground-truth bounding box larger than a specific threshold (IoU>0.5), are assigned with the corresponding ground-truth expression label cls (positive samples), while other RoIs are assigned as background (negative samples). To keep the training images balanced, we randomly sampled negative RoIs from labeled background RoIs. The total number of positive and negative RoIs should not exceed $N = 128$ for computing platform reasons. These positive and selected negative RoIs are formed into a minibatch for stochastic gradient descent (SGD). We adopt the same loss function as Faster R-CNN (Ren et al., 2017). For data augmentation purposes, we flipped all images in the horizontal direction in our infant expression training dataset.

3.2 Stabilization

Experiments have shown that a considerable number of false classifications occur between two correctly classified frames. Therefore, a Hidden Markov Model (HMM) for modeling the dynamics of expression changes in a video sequence is utilized, to reduce false positives and enhance the monitoring stability. In our application, we model seven expressions of interest and the background as states of the HMM. Because of limited video data availability with infants, our HMM is trained separately from the CNN expression classifier. The transition probability between two states is obtained by analyzing training sequences with the ground truth of states. The expression estimation updated by the HMM is calculated with a forward method, which is computed in two steps: prediction and update. In this work, we only use and discuss below a first-order HMM, meaning that the actual state only depends on the previous frame state.

Prediction. Given the observation sequence $O = \{o_1, ..., o_t\}$, the purpose of detecting and classifying the expression of a frame at time t (abbreviated as frame t) is to find the maximum a-posteriori probability of the state $p(q_t|o_t)$. In our problem, o_t corresponds to the union of all proposed RoIs in each frame, denoted as $o_t = \{b_{1,t}, ..., b_{n,t}\}$ with bounding boxes $b_{v,t}$, where n is the total number of RoIs in frame t and d that number for frame $t-1$. For each individual RoI with index u denoted as $b_{u,t-1}$ in frame $t-1$ $(1 \leq u \leq d)$, the posterior state probability of that RoI for frame t can be estimated by

$$p(q_t|b_{u,t-1}) = \sum_{i=1}^{k} p(q_t|q_{t-1})p(q_{t-1}|b_{u,t-1}). \qquad (1)$$

Here, k denotes the total number of states. This probability should be computed for all RoIs in frame $t-1$.

Update. For each RoI $b_{v,t}$ in frame t, where $1 \leq$

$v \leq n$, the posterior probability $p(q_t|b_{v,t})$ is found by

$$p(q_t|b_{v,t}) = \frac{1}{Z}\frac{1}{m}\sum_{v=1}^{m} Ov(v,u)p(b_{v,t}|q_t)p(q_t|b_{u,t-1}), \tag{2}$$

where the binary overlap function $Ov(v,u)$ indicates when there is an IoU overlap of RoI $b_{u,t-1}$ and $b_{v,t}$ of more than at least 70% or more. In case there is no overlap, the contribution becomes zero, because $Ov(v,u) = 0$.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

Public datasets containing only infants are rare, and datasets with ground-truth annotations for infant expressions are virtually not available. To our best knowledge, only two public datasets exist (Zamzmi et al., 2018) for infant pain/discomfort analysis, which are *COPE/iCOPE* collected (Brahnam et al., 2006) and another dataset collected from *YouTube videos* described in (Harrison et al., 2014). However, none of these two datasets can be used for our purpose. To solve this, we have manually collected a dataset consisting of $16,165$ infant images from the Internet, and expressions are manually labeled by bounding boxes. After the dataset augmentation, the total number of images used for training achieves $32,330$. Moreover, to evaluate the infant monitoring system both clinically and practically, we have first randomly selected 11 from 45 video sequences of different infants up to 2 years old with challenging situations, such as large head-pose deviation and object occlusions, which were denoted as *Clinic* for validating the clinical usage. The rest of the video sequences is used to train the HMM. All these videos were recorded at the Maxima Medical Center (MMC), Veldhoven, the Netherlands, satisfying the ethical standards of the institution, allowing usage after obtaining a written consent from the parents. Videos for infant discomfort expression are captured when experiencing pain from a heel prick, placing an intravenous line, or a vaccination, whereas other expressions are captured when infants stay at the hospital for medical care. All selected videos last at least 2 minutes, which is the required time for professional observation and pain scoring. Besides this dataset, similar to (Harrison et al., 2014), we have also collected 72 video sequences from *Youtube* denoted as *Youtube*. The purpose of this dataset is to validate for a practical application and for benchmarking with state-of-art methods. Here, each video sequence contains one different infant. All videos were uploaded by *Youtube* users, therefore video qualities are quite different (varying between good and poor) because of variable conditions. Moreover, these video sequences also contain certain frames without infants, which can be utilized to evaluate the reliability of all methods. Videos in both testing datasets are fully annotated. The method of annotating the ground truth of each frame is identical to images in the training dataset. All ground-truth annotations of both train and test sets are given by professional practitioners at the MMC. The number of frames for each expression in both test datasets are depicted in Table 1. As shown, the number of frames in each dataset exceeds 10,000 images, which is much larger than experimental data of the previous work (Sun et al., 2018; Zhi et al., 2018).

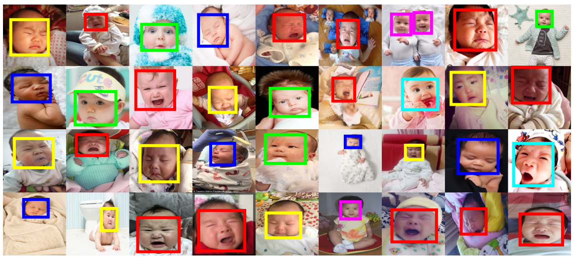


Figure 2: Examples of infants collected from Internet in the training dataset. Different color bounding box annotates each corresponding expression.

4.2 Metrics

In order to evaluate the accuracy of infant expression detection, the VOC2007 metrics (Everingham et al., 2010) are utilized in all the evaluational experiments, since VOC2007 is specifically designed for detection and classification tasks. By definition, the Average Precision (AP) indicates the detection accuracy of each expression, while the mean Average Precision (mAP) shows the overall performance of seven expressions of interest. For analyzing the consistency of the system output, the frequency of the False Expression Changes (FEC) is computed for each dataset. A false expression exchange is defined as an occurrence of an expression change between one frame to the next neighboring frame, that does not correspond to the exchange indicated by the ground truth. This can happen due to the false expression classification and the localization of bounding boxes with false detections. In addition to computing the frequency of the FEC, a second metrics is introduced, defined as the reduction rate, for analyzing the temporal stability. This analysis applies to the system only using R-CNNs and our proposed with temporal consistency

Table 1: Total number of frames within video sequences showing respective expressions in both test datasets.

	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	All
Clinic	3,523	7,940	2,587	166	1,462	1,246	1,399	18,323
Youtube	3,499	4,622	631	2,592	1,046	1,185	821	14,396

processing. The reduction rate is defined as:

$$R = \frac{F - F^{'}}{F}, \tag{3}$$

where $F^{'}$ represents the frequency of occurrence of false changes when temporal analysis is used, and F denotes the count of false expression changes without temporal processing.

4.3 Experimental Results

This section first provides results regarding the expression detection robustness using VOC2007 metrics (Everingham et al., 2010). Then, the consistency analysis of the system output is evaluated with the metrics discussed above.

A. Accuracy Evaluation. The performance of discomfort detection achieved by our multi-class network is compared with conventional methods, such as (Li et al., 2016) (Sun et al., 2018), which adopted HOG and LBP features as facial descriptors. Tables 2 and 3 present the experimental results of multi-class classification, based on CNN features evaluated with *Clinical* and *Youtube* datasets. The scores clearly indicate that the conventional methods using descriptors such as HOG and LBP, can hardly distinguish subtle facial expressions, such as Joy, Open mouth and Unhappy. Moreover, these methods are only capable of detecting Discomfort and Neutral with a low precision, which is not sufficient for implementation in an infant monitoring system. In contrast, the proposed CNN-based framework achieves an attractive score for the mAP of 82.3% and 83.4% for overall performance in Tables 2 and 3, respectively. For discomfort detection, it obtains an AP of 89.0% evaluated with *Clinic*, and an AP of 90.3% with the *Youtube* set. It can be observed that the accuracy and robustness of discomfort detection using CNN architectures are significantly increased compared with the conventional methods.

When combining an HMM model with the detection, the overall system outperforms solely Faster R-CNN (Ren et al., 2017) by 4.9% and 0.1% with the *Clinic* and *Youtube* sets, respectively. Here, it can be noticed that an increase in the AP occurs for Discomfort by 1% and 0.3% for the same datasets, respectively, compared to only using Faster R-CNN (Ren et al., 2017). This gain occurs because the trained R-CNN classifier performs less accurate in ambiguous expressions between Unhappy and Joy, and the combination of Discomfort and Unhappy. However, by adopting a dynamic model (HMM), the performances for these expressions are enhanced, due to the stabilizing information of the previous frame. In addition, the benefit of using HMM for *Youtube* is limited compared to that for *Clinic*. This is explained by the duration of video sequences in *Youtube* being shorter (a few seconds) and consistent expressions occurring in each sequence. Therefore, it is difficult for HMM to improve the performance further. Fig. 3 shows examples of the expression detection obtained by the proposed method evaluated with *Clinic*.

B. Output Consistency Evaluation. To improve the accuracy of expression detection for infant monitoring, the consistency of the proposed system is evaluated when using the HMM compared to solely using R-CNNs. Fig. 4 shows a bar plot of the false changes between expressions (Discomfort, Unhappy and others) occurring in both testing datasets. Discomfort and Unhappy are mainly taken into consideration for this experiment, because these two expressions carry more clinical information for facilitating a diagnosis. It can be seen from Fig. 4 that the false exchanges are significantly reduced, especially between Unhappy and other expressions. Furthermore, Table 4 provides the reduction ratios of FEC within Discomfort, Unhappy and others. It can be noticed that the FECs are overall reduced, while the reduction rate is preserved up to 65%. Specifically, the false changes between Discomfort and other expressions, which is the main reason of causing false chirps for an infant monitoring system, are significantly reduced by 64.9% and 35.2% for *Clinic* and *Youtube*, respectively. Fig 5 portrays an exemplary sequence of expression detections with and without the HMM. Comparing the two figures, it can be readily seen that after using the HMM, the detection becomes less noisy, resulting into a more stable and reliable performance.

C. Discussion: Although a high consistency is obtained by using an HMM, certain drawbacks are still noticeable. For example, the detection score of each class updated by an HMM highly depends on the performance of the trained classifiers. When HMM is combined with a less accurate classifier, the detection will be stuck at false positives. However, by jointly training Faster R-CNN and HMM in an end-to-end fashion, these false positives can be further reduced.

Table 2: Average precision for 7 expressions and the corresponding mAP for each method evaluated with *Clinic*. Boldface numbers indicate the highest score.

	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	mAP
LBP+SVM	0.349	0.245	0.341	0.007	0.017	0.068	0.045	0.153
HOG+SVM	0.254	0.154	0.608	0.077	0.095	0.026	0.033	0.178
F. R-CNN	0.880	0.838	0.873	0.578	0.811	0.534	0.907	0.774
F. R-CNN + HMM	**0.890**	**0.870**	**0.883**	**0.728**	**0.854**	**0.633**	**0.906**	**0.823**

Table 3: Average precision for 7 expressions and the corresponding mAP for for each method evaluated with *Youtube*. Boldface numbers indicate the highest score.

	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	mAP
LBP+SVM	0.201	0.513	0.091	0.433	0.037	0.054	0.001	0.190
HOG+SVM	0.129	0.377	0.091	0.466	0.162	0.048	0.045	0.188
F. R-CNN	0.900	0.752	0.969	**0.846**	**0.720**	**0.742**	**0.904**	0.833
F. R-CNN + HMM	**0.903**	**0.783**	**0.983**	0.831	0.608	0.713	0.902	**0.834**

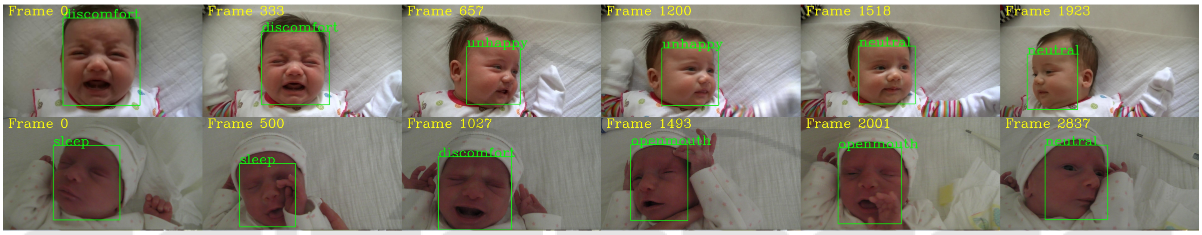
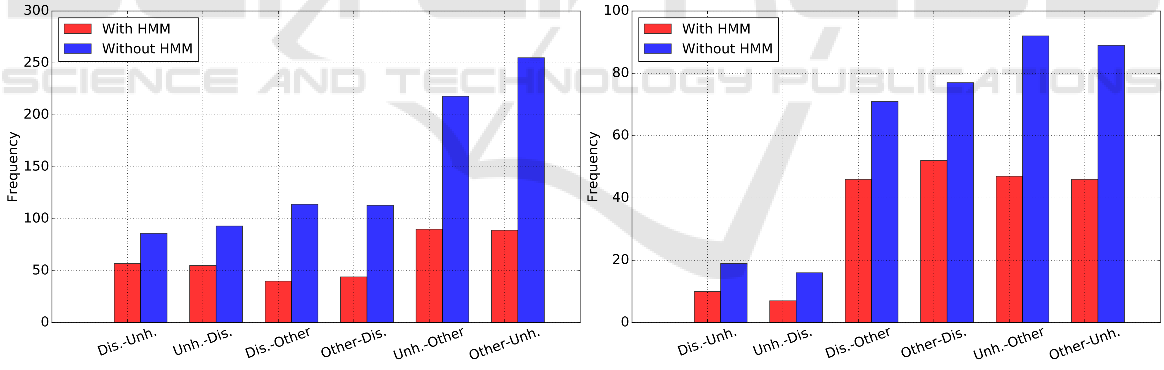


Figure 3: Examples of expression detection obtained by the proposed algorithm evaluated with Clinic.



(a) (b)

Figure 4: Frequency of the false expression changes (FECs) between the expressions of interest. (a) *Clinic*; (b) *Youtube*.

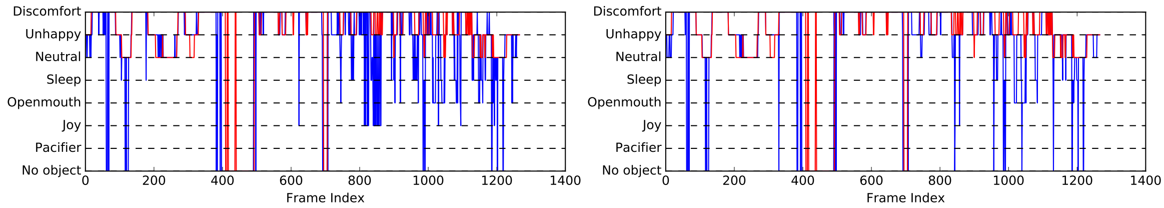


Figure 5: Examples of expression detection for an exemplary sequence. Left: detection by Faster R-CNN without HMM, right: detection with HMM. Red lines represent ground-truth information, blue lines denote detection by proposed algorithms.

Table 4: Reduction rate of false expression changes (FECs) between frames of R-CNN framework with the adoption of HMM compared to solely Faster R-CNN. Boldface numbers indicate the highest score.

	Clinic	*Youtube*
Discomfort - Unhappy	0.337	0.474
Unhhappy - Discomfort	0.409	**0.563**
Discomfort - Other	0.649	0.352
Other - Discomfort	0.610	0.325
Unhappy - Other	0.587	0.489
Other - Unhappy	**0.651**	0.483

In our experiment, the algorithm is executed on a GTX-1080ti GPU, which achieves a frame rate of 7 fps. It can be assumed that with a more advanced GPU, or combining a tracking method, the computation speed will increase, and therefore allows model usage in a real-time infant monitoring system.

5 CONCLUSIONS AND FUTURE WORK

This paper has proposed a near real-time video-based infant monitoring system, using Faster R-CNN combined with a Hidden Markov Model. The HMM increases the stability of decision making over time and reduces the noise in expressions. Differentiating from the conventional methods applying face detection and expression classification separately, we have trained a ConvNet detector that directly outputs the expressions. The experimental results have shown an AP achieving up to 90.3% for discomfort detection, which provides a dramatic accuracy increase compared to conventional methods (larger than 50%). The high-accuracy discomfort detection can be combined with some disease analysis such as GERD. In addition, the consistency of the system output is evaluated, and the experimental results have shown that the false expression changes between frames can be significantly reduced with a temporal analysis. In the future, as more video sequences of infants become available, we will train our expression classifier and temporal analysis end-to-end.

REFERENCES

Brahnam, S., Chuang, C.-F., Shih, F. Y., and Slack, M. R. (2006). Svm classification of neonatal facial images of pain. In Bloch, I., Petrosino, A., and Tettamanzi, A. G. B., editors, *Fuzzy Logic and Applications*, pages 121–128.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2017). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 39(4):677–691.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Fotiadou, E., Zinger, S., Tjon a Ten, W. E., Oetomo, S., and de With, P. H. N. (2014). Video-based facial discomfort analysis for infants. In *Proc. SPIE 9029, Visual Information Processing and Communication V, 90290F*.

Harrison, D., Sampson, M., Reszel, J., Abdulla, K., Barrowman, N., Cumber, J., Li, C., Nocholls, S., and Pound, C. M. (2014). Too many crying babies: a systematic review of pain management of practices during immunizations on youtube. *BMC Pediatrics*, 14:134.

Hicks, C., Baeyer, C., Spafford, P., van Korlaar, I., and Goodenough, B. (2001). The faces pain scale - revised: Toward a common metric in pediatric pain measurement. *Pain*, 93:173–83.

Jaskowski, S. K. (1998). The flacc: A behavioral scale for scoring postoperative pain in young children. *AACN Nursing Scan in Critical Care*, 8:16.

Li, C., Zinger, S., Tjon a Ten, W. E., and de With, P. H. N. (2016). Video-based discomfort detection for infants using a constrained local model. In *2016 Int. Conf. Systems, Signals and Image Proc. (IWSSIP)*, pages 1–4.

Lin, F., Hong, R., Zhou, W., and Li, H. (2018). Facial expression recognition with data augmentation and compact feature learning. In *2018 25th IEEE Int. Conf. on Image Processing (ICIP)*, pages 1957–1961.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conf. on Comput. Vis. Patt. Recog. (CVPR)*, pages 3431–3440.

Mueller, M., Smith, N., and Ghanem, B. (2017). Context-aware correlation filter tracking. In *2017 IEEE Conf. Comp. Vision Pattern Recogn. (CVPR)*, pages 1387–1395.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.

Sullivan, M. and Lewis, M. (2003). Emotional expressions of young infants and children. *Infants & Young Children*, 16:120–142.

Sun, Y., Shan, C., Tan, T., Long, X., Pourtaherian, A., Zinger, S., and de With, P. H. N. (2018). Video-based discomfort detection for infants. *Machine Vision and Applications*.

Tavakolian, M. and Hadid, A. (2018). Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition. *2018 25th IEEE Int. Conf. on Image Proc. (ICIP)*, pages 1952–1956.

Witt, N., Coynor, S., Edwards, C., and Bradshaw, H. (2016). A guide to pain assessment and management in the neonate. *Current emergency and hospital medicine reports*, 4:1–10.

Zamzmi, G., Kasturi, R., Goldgof, D., Zhi, R., Ashmeade, T., and Sun, Y. (2018). A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Reviews in Biomedical Engineering*, 11:77–96.

Zhi, R., Zamzmi, G., Goldgof, D., Ashmeade, T., and Sun, Y. (2018). Automatic infants' pain assessment by dynamic facial representation: Effect of profile view, gestational age, gender, and race. *Journal of clinical medicine*.