Infant Facial Expression Analysis: Towards a Real-Time Video Monitoring System Using R-CNN and HMM

Cheng Li , A. Pourtaherian , L. van Onzenoort, W. E. Tjon a Ten, and P. H. N. de With , *Fellow, IEEE*

Abstract—The manual monitoring of young infants suffering from diseases like reflux is significant, since infants can hardly articulate their feelings. In this work, we propose a video-based infant monitoring system for the analysis of infant expressions and states, approaching real-time performance. The expressions of interest consist of discomfort, unhappy, joy and neutral, whereas states include sleep, pacifier and open mouth. Benefiting from the expression analysis, the discomfort moments can also be used and correlated with a symptom-related disease, such as a reflux measurement for the diagnosis of gastroesophageal reflux. The system consists of three components: infant expressions and states detection, object tracking and detection compensation. The proposed system is based on combining expression detection using Fast R-CNN with a compensated detection using analyzing information from the previous frame and utilizing a Hidden Markov Model. The experimental results show a mean average precision of 81.9% and 84.8% for 4 infant expressions and 3 states evaluated with both clinical and daily datasets. Meanwhile, the average precision for discomfort detection achieves up to 90%.

Index Terms—Fast R-CNN, HMM, infant monitoring, near real-time application.

I. 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. As a result, caregivers or parents can only estimate the needs of a young infant by analyzing their expressions, body movements and sound. Infant expressions are informative and can convey signals describing their mood [1] and sometimes symptoms. For example, a discomfort expression will give a possible indication of particular symptom-related disease, such as reflux (some acid reflux will cause pain), which

is common for young infants. Alternatively, a joyful expression will imply pleasure and interest on certain events perceived by infants [1]. Over the years, systems have been developed for deciphering facial expressions of infants based on subjective descriptors. However, these systems require practitioners to observe infants for a certain amount of time, which is laborious and time-consuming. Moreover, a continuous monitoring and nursing of an infant provided by professionals would be too expensive in practice, when such care would be required by an infant. To address this problem, an automated video-based infant monitoring system can be implemented by analyzing expressions as an auxiliary assessment tool. This approach would differentiate from other methods, such as cable-based heart-rate monitoring, since video-based expression analysis has a non-interventional character. Besides, the detected discomfort moments can be correlated with a symptom-related reflux measurement for gastroesophageal reflux disease (GERD) diagnosis. By analyzing the discomfort moments and the reflux moments, the symptom-related reflux can be searched automatically from the numerous reflux moments. As a result, this approach allows the GERD diagnosis to be very efficient and accurate, which is appealing and welcomed by pediatricians. However, to design such a system, several challenges have to be resolved.

First, the system should perform face detection, even when the infant has large head poses deviating from the camera view. Second, facial expressions of infants are significantly different from those of adults. Hence, some of the state-of-the-art face detection and expression analysis methods trained with adult datasets mostly fail in pediatric applications. Therefore, a specific approach is needed for infants. Third, public infant datasets are rare in contents, therefore, training a robust classifier with a large number of parameters becomes intractable. Last but not least, the trained classifier should be able to detect expressions when faces are partially occluded by objects. To overcome the aforementioned challenges, we propose to use Convolutional Neutral Networks (CNNs) because they have proven successful in multiple computer vision tasks with a noticeable performance, like object detection and facial expression recognition [2]–[4] and they can be robust when properly trained. From these CNNs, we consider to construct a video-based framework by combining Fast R-CNN [5] and a tracking method, which is not only robust to challenging situations as aforementioned, but increases the executional speed. Fast R-CNN is chosen for two reasons. First, Fast R-CNN has been demonstrated to give
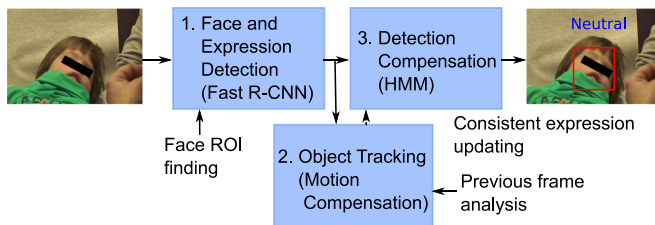
Fig. 1. Main processing stages of the proposed method for the infant monitoring system. The text in brackets indicates the techniques. Markov modeling over time improves the time consistency of the detection of expressions.

state-of-the-art performance for object detection [5]. Second, since temporal information is available from video analysis, an exhaustive search over each entire frame (Faster R-CNN [6]) will become clearly computationally expensive without offering much extra performance. For these reasons, Fast R-CNN enables balancing the computational costs and the accuracy for expression detection. Finally, with the adoption of a dynamic model to estimate the temporal expression changes, the output of expression detection becomes more stable and reliable, which is novel to this application. The block diagram of the proposed infant monitoring system is shown in Figure 1.

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

1) Different from the traditional discomfort analysis for infants [7]–[9], which explores face detection and discomfort analysis, we train our CNNs-based framework to find Regions of Interest (ROIs) which are directly oriented to expression analysis without extra effort for specific face detection.

2) Since infant expressions are subtle and instinctive, we distinguish infant expressions into four expressions and three states, rather than a binary discomfort detection.

3) A dataset consisting of more than 10,000 young infant images has been manually collected for training the CNN-based methods, while the expression annotations are defined by pediatricians. This infant dataset will be uploaded and made public to researchers shortly. The experimental results have shown that the dataset is large and diverse enough to successfully train CNN-based methods, thereby making the proposed dataset of significant importance for the infant monitoring task.

4) To the best of our knowledge, Fast R-CNN combined with tracking and HMM is for the first time adopted for this application. In addition, the executional speed is increased by using the tracking techniques, while the reliability for expression detection on challenging situations are improved by HMM. The experimental results have shown that the performance of discomfort detection based on our framework is significantly enhanced compared to state-of-the-art methods using conventional techniques, and also outperforms other methods based on CNNs.

The remainder of this paper is organized as follows. Section II discussed the difficulties and requirements of designing an automated infant monitoring system. Section III briefly introduces

some related work in both clinical and technical domain. Afterwards, Section IV describes our definition of infant expressions and states. Next, the details for the system design are explained in Section V. The experimental results for infant expression analysis are provided in Section VI. Finally, Section VII presents conclusions.

II. PROBLEM STATEMENT

Human expressions are usually subtle and mixed and very ambiguous to distinguish, especially for situations, such as changes from one expression to another. Furthermore, for a surveillance and infant monitoring task, expression analysis becomes even more challenging due to practical difficulties, such as having less facial key features within infant faces, extreme light conditions and occlusions blocking the visibility of the face [10]. Lack of facial features is normally caused by a non-frontal projection from an infant face to the camera, when an infant has large pitch and yaw angles deviating from the camera view. In this case, some key facial features for identifying discomfort, such as a wide-opened mouth and one-side nasolabial winkles, will be absent in the obtained 2D image. Meanwhile, extreme light conditions usually occur due to infant sleeping habits, so that they require a very dark environment for sleeping during the daytime. Therefore, to design a robust infant monitoring system that can be employed in a hospital environment, several requirements should be satisfied.

- Preferably, the system should not be invasive and introduce no extra pain to infant patients.
- The accuracy of infant expression detection should be as high as possible even when aforementioned practical difficulties are present. Nevertheless, the system should emphasize discomfort detection against other expressions, since infants normally require an immediate care when they feel discomfort/pain.
- The system should have real-time performance, so that it can handle and signal the actual reality of the physical state of infant patients.
- A continuous output of the system is required, especially when false detections of discomfort occur in between two adjacent frames containing neutral expressions should be prevented. Because such false positives are inconvenient for the care-givers and therefore should be reduced.

To meet the aforementioned requirements, we propose an infant monitoring system based on video analysis. With this concept, video sequences of multiple infants are used to train the framework, after which the expression detection can be applied to each frame of new video sequences. The results and detections are indicated and stored. In this paper, we propose methods and techniques specifically dealing with the above requirements, while aiming at an experimental implementation of such a system into a clinical environment for further validation.

III. RELATED WORK

Because of the limited abilities in verbal communication and associative thinking of infants, the absence of regular pain

assessment makes pain for infants normally under treated. However, 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 objective measurement of pain.

A thorough review of pain assessment tools used in the clinic for neonates is provided in [11]. Among these proposed tools, facial expressions are mainly used as indicators of measuring pain, such as for the revised FACES pain scale [12], FLACC [13], the Wong-Baker Faces scale, the neonatal facial coding system (NFCS) [11] and the 10-cm visual analog scale. Due to the simplicity of implementation, these tools are widely used in many healthcare settings to assess the pain of a pediatric patient. The revised FACES pain scale, FLACC, and the Wong-Baker Faces scale reveals the levels of pain intensity by different facial expressions, where each level is scored by a numerical number. Contrary to these tools using full faces, NFCS deciphers pain with facial actions, such as lower brows, squeezed eyes, and deepened nasolabial furrows, etc. The ultimate pain score is therefore obtained by accumulating the occurrence of the facial actions. In addition to pain assessment, researchers also consider to explain and describe different infant expressions. For example, Sullivan and Lewis [1] distinguish infant expressions as surprise, enjoyment, physical pain, etc.

Inspired by these pain assessments used in clinical settings, some automated pain detection systems for neonates have been studied based on facial expression analysis. In [7], a semi-automated system is proposed for discomfort detection, which adopts an Active Appearance Model (AAM) for facial appearance modeling. However, facial contours in the key frames of a video sequence are required to be labeled manually, which hampers the possibility for a real-time application. Zhi *et al.* [14] proposed an automatic infants' pain detection by using geometric features. After that, a Support Vector Machine (SVM) is utilized to distinguish pain from no pain. In [8], it is proposed to exploit an automated classification model, based on appearance features extracted by local binary patterns. Sun *et al.* [9] 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. Sikka *et al.* [15] discussed an automated assessment of pain for postoperative children using a logistic regression. The author in [16] proposed a method to fuse a set of features extracted by SIFT descriptors applied on a full image and the corresponding subregions. After that, pain expression is obtained by a SVM classifier. Besides this, the author in [17] proposed a multi-modal pain analysis approach for infants based on extracting strain features with Optical Flow.

All these presented methods adopt the processing chain of face detection, landmark localization and discomfort detection based on conventional classifiers. Therefore any processing component can lead to a failure of the final expression analysis. In these processing steps, landmark localization is particularly considered as the most critical part, since facial features are extracted relying on aligned fiducial facial points. However, even with a state-of-the-art method [18], the landmark localization will fail when a head orientation significantly deviates from the frontal view, which inevitably brings negative impact on expression analysis. Moreover, all these works model infant expressions as a binary situation, such as pain and no pain, or discomfort and comfort. However, infant expressions are more subtle and complicated, and these binary classifiers will show false positives when a joy expression occurs. Recently, CNNs become prevalent for complicated texture representations, and have been widely used in many computer vision tasks. Tavakolian and Hadid [19] proposed a pain expression intensity estimation based on CNNs and a binary coding. Lin *et al.* [20] have proposed a CNN-based expression classifier trained with data augmentation for adults. Li *et al.* [4] proposed a facial expression recognition with Faster R-CNN for adults, which is similar to our work. However, all these CNNs-based methods for expressions can not be applied directly for infants, since expressions of interest for adults, such as surprise and anger can hardly appear on an infant face. Hammal *et al.* [21] apply CNNs for action unit detection for infants. However, this approach requires the face of infants to be a frontal view, so that the landmark localization can be accurate. As an alternative to video-based infant monitoring, researchers have also exploited methods using other features to indicate pain and discomfort, such as vocal features extracted from infant cries [22], [23]. To increase the robustness to challenging situations, in this paper, we will propose an infant video-based expression analysis algorithm using Fast R-CNN and a dynamic model, while aiming at implementing a near real-time automated infant monitoring system.
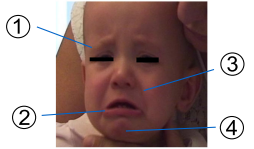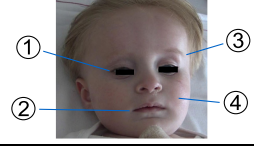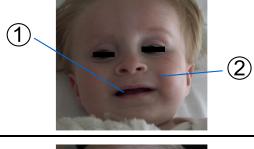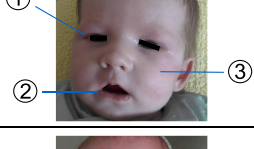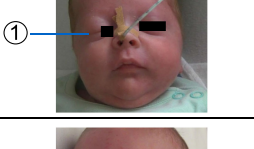
IV. INFANT EXPRESSION AND STATE DEFINITION

Since the main task of an automated infant monitoring system is to detect pain, we consider pain and discomfort detection as the first priority of our investigation. According to the conventional methods for pain analysis, infant expressions are classified as a binary situation (pain and no-pain). However, this strategy is limited because the accuracy of pain/discomfort detection cannot be further improved by increasing the amount of training images to reach a high performance. The reason is that infant expressions are subtle and cannot be adequately classified by a binary classifier. Therefore, to address this complication, a more detailed expression division is required.

Based on the work in [1], [24], we classify infant expressions and states into seven categories. Expressions include discomfort/pain, unhappiness, neutral and joy, whereas states consist of open mouth, sleep and pacifier. These expressions and states are defined, since they either appear frequently or are present on infant faces for a long time, and their existence makes it very difficult for a classifier to distinguish them from discomfort. For example, a discomfort, joy and unhappy face all have deepened nasolabial folds as their Facial Action Unit (FAU). As a result, in this section, we will discuss the definition of each expression and state of interest, as well as their corresponding FAUs. Moreover, the situations when these expressions occur will also be briefly introduced. Table I portrays visual examples of each expression and state, as well as their FAUs. The detailed explanation of the

TABLE I
EXPRESSIONS AND STATES DEFINITION AND THEIR CORRESPONDING FAUs. (EYES OF THE INFANTS ARE OVERLAID WITH BARS TO PRESERVE ANONYMITY.)

Expression and States	Cases of Occurrence	Exemplar Images	Facial Action Units
1. Discomfort/Pain	Pain expression occurs when infant receives tissue damage during standard pediatric procedures (eg, heel lance, vaccination) [1]. Discomfort expression usually occurs when infants feel hunger, stressed or less secure.		① - Mid-brow bulge ② - Deepened nasolabial furrow ③ - Vertical stretched mouth ④ - Eye squeezing ⑤ - Chin quivering
2. Unhappy	The occurrence of this negative expression reflects general distress or unhappiness. According to our observations, this expression normally occurs as a transitional state from a positive to a negative expression, or vice versa.		① - Lowered brow ② - Down-turned mouth corner ③ - Prominent nasolabial fold ④ - Raised chin
3. Neutral	Neutral faces normally appear when infants stay in a quiet and awake state.		① - Opened eye ② - Closed mouth ③ - Relaxed brow ④ - Relaxed cheek
4. Joy	This expression includes two behaviors, smiling and laughter. Laughter mainly appears in awake and alert infants [1].		① - Widened mouth with corners raised ② - Deepened nasolabial furrow
5. Open mouth	Open mouth normally conveys less information as other expressions, therefore it is categorized as a state. It belongs to a family of non-expression, but is distinct from a neutral and awake face.		① - Opened eye ② - Opened mouth ③ - Relaxed cheek
6. Sleep	Sleep is better defined as a state than an expression, and is the most important behavior of infants. During infant sleep, they normally stay in a calm and quiet state. This behavior occurs for infants during the most of the day, while only afternoon and nighttime for toddlers.		① - Closed eye
7. Pacifier	Pacifier is not an expression, therefore it is categorized as a state. This occurs when infants are calm and quiet. In addition, the pacifier can be used at anytime and at any mode (asleep or awake).		① - Pacifier

FAUs, expressions and states in the table are self-explaining, and accurate enough to define the corresponding models for computing.

V. METHOD

In this section, the methods specifically designed for constructing the infant monitoring system will be provided. The framework of the system can be divided into three stages, direct expression detection, object tracking and detection compensation (see Figure 1). In our system, expression detections are realized by Fast R-CNN and VGG-net [25], in which the object region proposals are obtained by the proposed tracking method. Nevertheless, in the first frame of a video sequence and when object tracking fails, the object region-of-interest proposals are generated by selective search [26]. Moreover, to reduce false positives and improve the reliability of the detected expression,

we apply a Hidden Markov Model (HMM) on the detection results of the current frame using the information from the previous frame. The details of each component is depicted in Figure 2. As a result, the system outputs a stable and consistent detection for infant expressions and states. In the following section, each stage is now explained in detail. This research is approved by Medisch Ethische Toetsing Commissie (METC) of Maxima Medical Center (MMC), Veldhoven, the Netherlands with the protocol number of 15.065.

A. Expression Detection Training

Due to the limitation of existing datasets on infants, it is impractical to train all parameters of VGG-Net from scratch. Besides, our application is specific for infant monitoring, and it is known that infant expressions are more instinctive than expressions of adults. Therefore, instead of training the VGG-Net
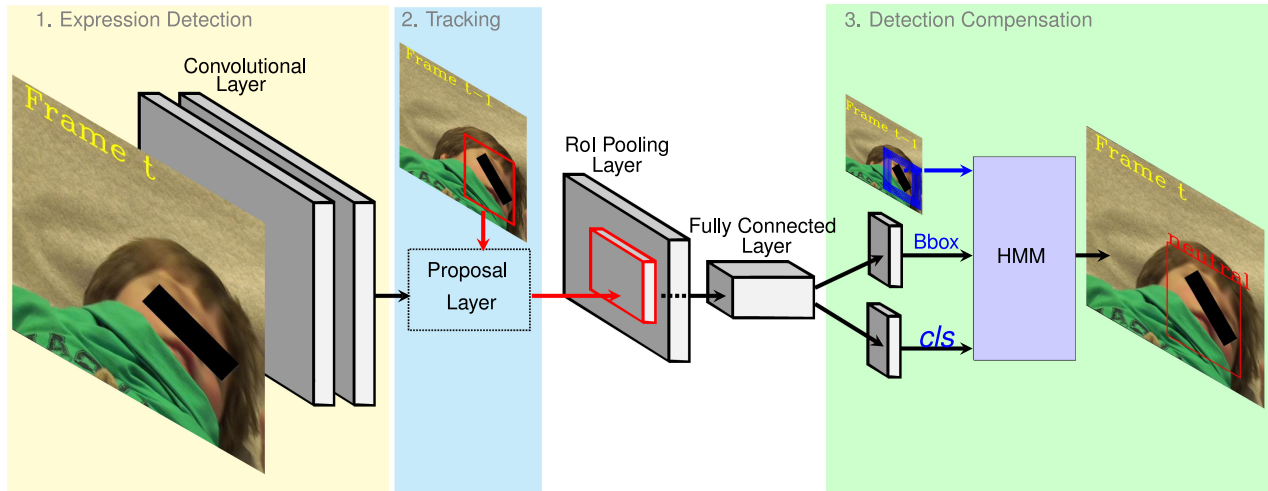
Fig. 2.   Flowchart of the system design, in which potential ROIs for the current frame are proposed by the previous frame and the detection of expression is updated based on a Hidden Markov Model (HMM) using the information of the previous frame. Red and blue bounding boxes in the image represent the detection and ROI proposals of the corresponding image, respectively. Bbox and cls are the abbreviations of bounding box and expression classes, respectively.

using public face datasets [27] which contain mostly adult faces, we firstly pre-train parameters with the ImageNet dataset [28]. Then the pre-trained VGG-Net is fine-tuned with a dataset composed of infant images only, which is collected from the Internet. This dataset consists of images exhibiting challenging situations, such as large head poses (profile of a face). Since the first few convolutional layers are trained to represent basic features, such as edges, corners, etc., we only update parameters in the last 9 convolutional layers. In order to map CNN features to classifications, we train three fully connected layers from scratch. The weights of the fully connected layers are initialized by sampling from a Gaussian distribution with zero mean and standard deviation of 0.01 for the first two layers and a standard deviation of 0.001 for the last layer. The last fully connected layer is adapted to 8 outputs, corresponding to the four expressions and three infant states of interest together with the background.

Similar to training for Fast R-CNN object detection [6], the input of the VGG-Net for training 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 and states of the corresponding bounding box g_i. In this application, cls corresponds to one of the following expression states, as also indicated in Section IV: Discomfort/pain, Unhappy, Neutral, Sleep, Joy, Open mouth and Pacifier.

Between the last convolutional layer and the first fully connected layer, we apply a proposal layer to indicate ROIs. For the training phase, ROI proposals are generated by using a selective search [26] method. For the test phase, ROIs are proposed based on our tracking results, which will be explained in the next section. In the training phase, the ROIs that have an intersection over union (IoU) with the ground-truth bounding box larger than a specific threshold, are assigned with the corresponding ground-truth expression label cls (positive samples), while other ROIs are assigned as background (negative samples). In our
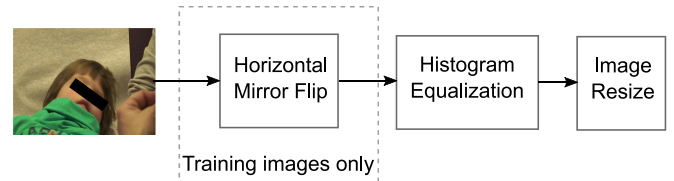

Fig. 3.   Flowchart of the pre-processing steps of the input images prior to the detection network.

experiment, we set the threshold for labeling a positive ROI as $IoU > 0.5$. In order 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 Fast R-CNN [6]. For data augmentation purposes, we flipped all images in the horizontal direction in our infant expression training dataset.

To address variations of the video input quality resulting from different datasets, some preprocessing of the images is initially applied prior to the expression detection. For the input image I_m, histogram equalization is applied to each signal channel for normalizing and increasing the signal contrast, after which it is spatially re-scaled to a standard image size. The procedure of pre-processing is shown in Figure 3. For optimization, we use stochastic gradient descent with the base learning rate equal to 0.001, gamma equal to 0.1, and momentum equal to 0.9. Both training and testing procedures are implemented using the Caffe [29] framework.

B. Object Tracking

In our application, an infant is under surveillance by monitoring its facial expressions in video sequences. Compared to

frame-based expression analysis, tracking infant expression in an interframe technique (video-based) is more efficient. Current state-of-art tracking methods [30] are designed for a general tracker, whose tracking template is updated on-line. However, in our application, our objective is to track expressions which are dynamic from frame to frame, so that updating templates on-line degrades the real-time performance. Since the expression classifier is trained off-line, we apply a tracking-by-detection manner that is suitable for our monitoring purpose.

For now, we are only interested in monitoring one infant at a time. Based on our observations of different infants in several video sequences, we have noticed that an infant behaves likely to be quiet in most of the video frames, such as sleeping, rather than being very active, such as moving the head constantly. Therefore, we assume that the spatial changes of the infant head are normally within a reasonable range. Under this assumption, ROIs of the current frame can be proposed, based on the detection in the previous frame. For the testing phase, we simply sample ROIs around the detection of the previous frame within a sliding window for the current frame. The center of the previous detection is considered as a reference, and we slide the window around the reference in both vertical and horizontal directions with a fixed interval. At each sliding-window center, an ROI is abstracted with the previous detection size. This ROI proposal method is similar to the anchor proposal in Faster R-CNN [6]. However, because of the consistency of head motion in our video sequences, the region proposal with one size ratio (the size of the previous detection) is sufficient to handle the movements towards or away from the camera. Compared to 300 region proposals in Faster R-CNN [6] and 2,000 region proposals in Fast R-CNN distributed within a whole frame, our tracking method generates on the average only 40 ROIs around the detection in the previous frame, which has been found sufficient for a reliable tracking. Compared to a frame-based expression analysis, such as Faster R-CNN [6], our video-based solution requires 7.5 times lower amount of ROIs for a comparable detection. This significant reduction facilitates the near real-time operation of our system.

C. Detection Compensation

In our testing phase, we have noticed that a considerable number of false detections occur between two correctly detected frames. Especially for frames with a transition state from one expression to another, the accuracy of this detection is low and unstable. This happens mainly due to the incapability of the trained CNN classifier to accurately distinguish these ambiguous expressions. Therefore, we utilize a Hidden Markov Model (HMM) for modeling the dynamics of expression changes in a video sequence, to reduce these types of false positives and enhance the stability of the monitoring. This procedure consists of two stages: prediction and update. In this section, we will explain the methodology in detail.

1) HMM Training: In our application, we model the four expressions and three infant states of interest together with the background as the hidden states of the HMM. Because of the limited availability of video sequences with infants, our HMM

is trained separately aside from the CNN expression classifier. The transition probability $p(q_t = s_j | q_{t-1} = s_i)$ is obtained by analyzing training sequences with ground-truth of states. The state-output probability $p(o_t | q_t)$ is approximated by the posterior probability of a state $p(q_t | o_t)$, which is based on Bayesian theory, specifying that

$$p(o_t|q_t) = \frac{p(q_t|o_t)p(o_t)}{p(q_t)}, \tag{1}$$

where $p(q_t)$ is the state probability, which can be estimated from the training set. Probability $p(q_t|o_t)$ is the softmax layer output of the VGG-Net classifier. In this calculation, we assume that all observations $\{o_1, \ldots, o_t\}$ follow an uniform distribution, therefore $p(o_t)$ is considered as a constant. 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 a first-order HMM, which means that the state estimation of the current frame only depends on the previous frame. The model state is stabilized by exploiting prediction and update cycles, which are discussed below.

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

$$p(q_t|b_{u,t-1}) = \sum_{i=1}^{k} p(q_t = s_j|q_{t-1} = s_i)p(q_{t-1} = s_i|b_{u,t-1}). \tag{2}$$

This probability should be computed for all ROIs with index u where $1 \leq u \leq n$.

3) Update: For each ROI $b_{v,t}$ in frame t with index v, the posterior probability $p(q_t|b_{v,t})$ is calculated by

$$p(q_t|b_{v,t}) = \frac{1}{Z}\frac{1}{m}\sum_{v=1}^{m} Ov(v,u)p(b_{v,t}|q_t)p(q_t|b_{u,t-1}), \tag{3}$$

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%. When there is no overlap, the contribution becomes zero as $Ov(v, u) = 0$. Here, Z denotes a normalization factor, and $p(b_{v,t}|q_t)$ is obtained by Eq. (1). An example of bounding boxes which are used to contribute to the posterior probability calculation for $b_{v,t}$ is shown in Figure 4. Finally, the adopted detection is the one with the highest posterior probability from all proposed ROIs in frame t.

VI. EXPERIMENTAL RESULTS

This section explains the used datasets to train and test our infant monitoring system. After that, the experimental results will be provided. Finally, we briefly address the computational complexity of the system.

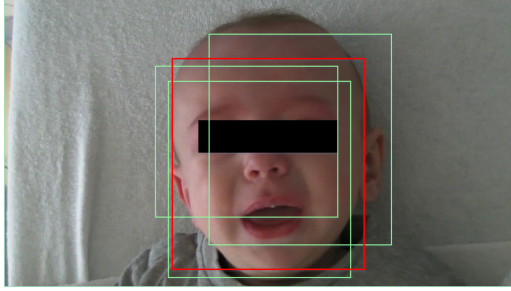|	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	All
Train	2,147	5,115	3,417	2,804	1,720	825	137	16,165

Fig. 4. Posterior probability of an ROI in frame t is calculated by averaging the contribution of bounding boxes in frame $t-1$, that have an IoU area overlap larger than a threshold of 70%. The red box represents the ROI in frame t denoted as $b_{v,t}$, and the green boxes represent bounding boxes ($b_{u,t-1}$) that have contributions to $b_{v,t}$.

A. Datasets

A sufficiently broad dataset is a key factor for training a reliable neural network. However, public datasets consist of 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 are available [31] for infant pain/discomfort analysis, which are *COPE/iCOPE* collected by Brahnam *et al.* [32] and another dataset collected from *YouTube videos* described in [33]. *COPE/iCOPE* contains 204 static images of cropped infant faces that are captured from 26 healthy infants. Since the proposed system is designed for long-time infant monitoring, the task focuses on detection applied to video sequences. As a result, such a dataset fails to meet the train and evaluation requirements. The dataset proposed in [33] consists of video sequences, however, the author annotated all video sequences with the FLACC [13] pain scale for pain assessment. Hence, instead of annotating for every single frame, a final pain score is given to each sequence, which is incompatible with our evaluation purpose.

Therefore, to train a VGG-Net classifier to be accurate and robust for expression analysis, we have manually collected 16,165 infant images from the Internet, and labeled the expression, based on the expression definition in Section IV by professional pediatrics in the MMC. All images selected for training contain no faces occluded by obstacles, except for images intended for training the Pacifier category. Therefore, all features for distinguishing expressions are clearly visible. Moreover, this training dataset contains varying head poses, such as semi-profile images. The ground-truth location of an infant face is also annotated with a bounding box. The majority of the young infants included in this training dataset are approximately under 2 years old. Besides, to increase the robustness for skin color, infants are selected from different continents, such as Asia, Europe and Africa. The detail of the number of training samples corresponding to each expression and state is shown in Table II.

It can be noticed that the training dataset is unbalanced, where the number of Pacifier images is far lower than other categories. This is because the texture of a pacifier is very distinct and representative, so that even with the current amount of training samples for Pacifier, the detection accuracy is comparable with other expressions and states. In addition, the amount of Unhappy is far lower than other expressions, which is due to practical difficulties for collecting images for this category (fewer images on the Internet).

To evaluate the infant monitoring system both clinically and practically, we have first randomly selected 11 video sequences of different infants with challenging situations, such as large head pose and object occlusions, which were denoted as Data-Clinic for validating the clinical usage. All these videos are recorded at a clinical hospital, observing the ethical standards of the institution allowing to use them 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 and states are captured when infants stay at the hospital seeking for medical care. All selected videos last at least 2 minutes, which is the time required for a professional to observe and give a pain score. Besides this data, similar to [33], we have also collected 67 video sequences from *Youtube* denoted as Data-Youtube. The purpose of this dataset is used as a validation for a practical application and a benchmark for comparing with the state-of-the-art methods. In this dataset, each video sequence contains one different infant. The *URL* of each video sequence is provided in the supplementary materials. All these videos are uploaded by *Youtube* users, therefore video qualities differ (from good to worse) due to the different cameras used for video acquisition. Moreover, these video sequences also contain certain frames which have no infant presenting, which can be utilized to evaluate the reliability of all methods.

Videos in both testing datasets are fully annotated. The ground truth of each frame is annotated according to facial actions depicted in Table I as well as the context of the video sequence. Since the testing videos are annotated by pediatricians and educated nurses for infants based on the aforementioned rules, these manual annotations therefore reflect the true status of the infants in the video sequences, i.e. the ground truth. For instance, when a video is filmed during a heel-prick procedure, an annotated Discomfort or Unhappy expression indicates a real discomfort or unhappy status. Note that for the output of the system, it only reflects the detected status of infants in the current frame. Therefore, only when the system detects an expression which is corresponding to the annotation of this frame, the output indicates a true interpretation of the infant status for the current moment.

However, when an infant face in the image is significantly occluded, such as missing the entire lower face (as shown in

TABLE III
TOTAL NUMBER OF FRAMES WITHIN VIDEO SEQUENCES SHOWING RESPECTIVE EXPRESSIONS AND STATES IN BOTH TEST DATASETS

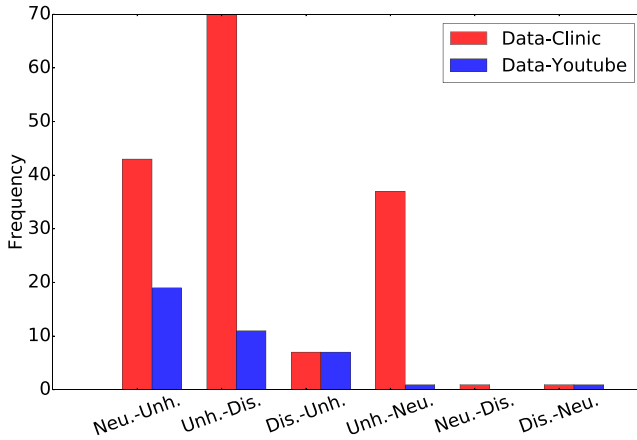|	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	All
Data-Clinic	3,523	7,940	2,587	166	1,462	1,246	1,399	18,323
Data-Youtube	2,863	4,300	631	2,583	1,026	1,169	821	13,393

Fig. 5.   Frequency of expression changes between Neutral, Unhappy and Discomfort in both testing datasets. (Neu. is the abbreviation for Neutral, while Unh. is for Unhappy, Dis. is for Discomfort).

Figure 7 the 1st and 7nd rows), it will be annotated as "no expression" presented. In this case, any detection in these images is considered as a false positive of the corresponding detected expression. If a face is partially occluded and most key-expression indicators are still visible, the ground-truth annotation will be determined by the visible indicators, the context of the video as well as the expressions of the previous and the coming frames. If it is still questionable to identify the expression with the aforementioned aspects, it will be annotated as "no expression" like in the training set.

The number of frames for each expression in both testing datasets are depicted in Table III. In addition, the numbers of expression changes between Discomfort, Unhappy and Neutral in both testing datasets are depicted in Figure 5. As shown in Table III, the amount of testing images in the Data-Clinic is unbalanced as well, and this is due to the fact that infants staying in the hospital are normally treated with pain killers for the pain control, which is compliant with the treatment protocol. Therefore, infants most likely show neutral expressions rather than discomfort and joy. Meanwhile, the images for Sleep are also unbalanced in Data-Youtube, this is because the difficulty of finding sleeping babies from *Youtube*, which leads to a far lower number of Sleep images compared to other expressions. Each frame of testing video sequences undergoes the same preprocessing procedures as the training images (shown in Figure 3) prior to the expression detection stage. The following sections will provide the experimental results of the proposed method.

B. Metrics

For evaluational metrics, the work in [7]–[9] uses a Receiver Operating Characteristic (ROC) curve. However, the ROC curve

normally presents the performance of a binary classifier. Instead and to acquire a fair comparison, VOC2007 metrics [34] are utilized in these evaluational experiments, since VOC2007 is specifically designed for detection tasks. By definition, an Average Precision (AP) indicates the detection accuracy of each expression, while the mean Average Precision (mAP) shows the overall performance of four expressions and three states of interest.

C. Results

In this section, all experiments are evaluated and compared between our proposed framework and three state-of-the-art methods. The first reference method is performing discomfort detection based on conventional techniques using Local Binary Patterns (LBP) and Support Vector Machine (SVM) for classification [7]. The second reference method is also exploiting conventional techniques, but now based on Histogram of Oriented Gradient (HOG) [9]. For enabling a fair comparison, a CNN face detector from Dlib [35] is added to the first two reference methods as a face detection step. The third reference method is using the Faster R-CNN network [36] for learning, which is potentially more powerful than the proposed method because it is more complex. For the Faster R-CNN method, the experiment is implemented by training the network with the same training dataset as used for our proposed method, and then being applied to each frame of the testing video sequences. The fourth method is our proposed method using Fast R-CNN with HMM modeling.

For evaluation, we will first provide the comparison of the four methods using binary outputs only (Discomfort and Neutral). After this, to understand the contribution of multi-expression detection for improving discomfort analysis compared to the binary case, the performance of each method for multi-expression detection will be presented.

Table IV shows the experimental results of discomfort detection obtained by binary detectors of each method, evaluated with both datasets. Despite the unbalance in the data due to the binary mapping, it can be observed that a binary detector based on a CNN network significantly outperforms the first two conventional methods for distinguishing discomfort by 50.6% and 71.7% for Data-Clinic and Data-Youtube, respectively. The proposed Fast R-CNN+HMM combination gives a similar performance as Faster R-CNN, despite the different complexity of the methods. The benefit of using the proposed combination becomes visible when employing the multi-expression detection in the following experiments.

Table V and Table VI present the experimental results of multi-expression detection for each method evaluated with two individual datasets. It is noticed from the results that the conventional methods can hardly distinguish subtle facial expressions,

TABLE IV
AVERAGE PRECISION FOR BINARY EXPRESSION DETECTION (DISCOMFORT AND NEUTRAL) AND THE CORRESPONDING mAP FOR EACH METHOD WITH THE TWO DATASETS, DATA-CLINIC AND DATA-YOUTUBE. BOLDFACE NUMBERS INDICATE THE HIGHEST SCORE

	Data-Clinic			Data-Youtube		
	Discomfort	Neutral	mAP	Discomfort	Neutral	mAP
CNN+LBP+SVM	0.352	0.721	0.536	0.168	0.609	0.389
CNN+HOG+SVM	0.258	0.644	0.451	0.091	0.605	0.348
Faster R-CNN	**0.858**	**0.898**	**0.878**	0.876	**0.887**	0.882
R-CNN + HMM	0.850	0.895	0.872	**0.885**	0.886	**0.886**

TABLE V
AVERAGE PRECISION FOR MULTI-EXPRESSION DETECTION (4 EXPRESSIONS AND 3 STATES) AS WELL AS THE CORRESPONDING mAP FOR EACH METHOD, EVALUATED WITH ONE DATASET, DATA-CLINIC. BOLDFACE NUMBERS INDICATE THE HIGHEST SCORE

	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	mAP
CNN+LBP+SVM	0.349	0.245	0.341	0.007	0.017	0.068	0.045	0.153
CNN+HOG+SVM	0.254	0.154	0.608	0.077	0.095	0.026	0.033	0.178
Faster R-CNN	0.880	**0.838**	0.873	0.578	0.811	0.534	**0.907**	0.774
R-CNN + HMM	**0.895**	0.802	**0.879**	**0.701**	**0.886**	**0.705**	0.868	**0.819**

TABLE VI
AVERAGE PRECISION FOR MULTI-EXPRESSION DETECTION (4 EXPRESSIONS AND 3 STATES) TOGETHER WITH THE CORRESPONDING mAP FOR EACH METHOD, EVALUATED WITH ONE DATASET, DATA-YOUTUBE. BOLDFACE NUMBERS INDICATE THE HIGHEST SCORE

	Discomfort	Neutral	Sleep	Joy	Open mouth	Unhappy	Pacifier	mAP
CNN+LBP+SVM	0.179	0.460	0.091	0.423	0.086	0.091	0.001	0.190
CNN+HOG+SVM	0.149	0.351	0.091	0.377	0.116	0.091	0.050	0.175
Faster R-CNN	0.891	0.830	0.983	0.845	**0.743**	**0.734**	0.904	0.847
R-CNN + HMM	**0.900**	**0.877**	0.983	0.845	0.738	0.685	**0.906**	**0.848**

such as Joy 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 81.9% and 84.8% for overall performance in Table V and Table VI, respectively. For discomfort detection, it achieves an AP of 89.5% evaluated with Data-Clinic, and an AP of 90.0% with the Data-Youtube set. From these experiments, we can conclude that the accuracy and robustness of discomfort detection using CNN architectures can significantly enhance the performance, compared to the conventional methods regardless of the number of expressions of interest.

When comparing the results of binary detection in Table IV with multi-expression detection in Table V and Table VI, the Fast R-CNN+HMM multi-expression system outperforms the same framework tuned to binary outputs by 4.5% and 1.5% for discomfort detection, evaluated with the Data-Clinic and Data-Youtube datasets, respectively. Apparently, for discomfort detection, a binary classification is insufficient for distinguishing subtle expressions and states, and therefore can hardly be improved further when the bottleneck is reached.

It can also be observed from Table V and Table VI that for multi-expression detection, the proposed Fast R-CNN+HMM outperforms Faster R-CNN for the overall performance by 4.5% and 0.1% with the Data-Clinic and Data-Youtube datasets, respectively. Especially for discomfort, the AP performance is increased by 1.5% and 0.9% for the same datasets. This gain occurs due to the fact that the trained network (regardless of Fast R-CNN and Faster R-CNN) is less accurate in distinguishing

ambiguous expressions between Unhappy and Joy, Discomfort and Unhappy. This may be caused by the limited number of Unhappy samples included in the training dataset. However, under the circumstances of lacking training images for specific expression categories, the performances for expressions and states can be partially enhanced (especially for Discomfort), by combining the derived information of the previous frame with that of the actual frame. This concept leads to a novel and cost-effective method extension in which HMM is used for obtaining a higher consistency of decision making. Consequently, the overall performance of the combination of Fast R-CNN and HMM is more stable and reliable. In addition, Figure 6 shows Precision-Recall Curves of discomfort detection obtained by all the evaluated methods with both datasets. As observed, the proposed method based on Fast R-CNN+HMM which is tuned as a multi-expression framework performs the best, while the Faster R-CNN with seven classes performs somewhat lower and the conventional methods using HOG and LBP perform worse. Examples of detection results obtained by our algorithm and Faster R-CNN [6] are shown in Figure 7.

Even though the CNN-based methods are mainly trained with images of infants below 2 years old, they can still detect some expressions and states (Discomfort, Neutral, Sleep, Open mouth, etc.) for toddlers (shown in Figure 7, the $1st$ and the 3rd row), which is beyond our expectation. This happens because the high-order features of expressions and states, such as Discomfort, Neutral and Pacifier *etc.*, represented by CNNs are less distinct for neonates and toddlers than assumed. Consequently, this brings a new possibility of designing a general infant monitoring system regardless of age and spices. However, for detecting other

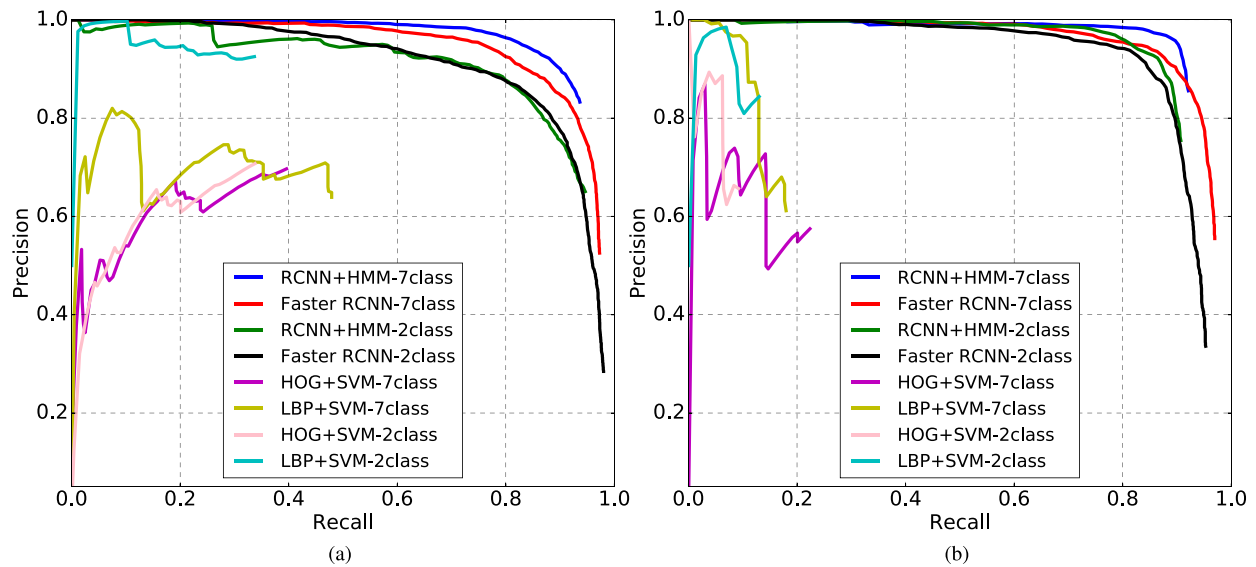Fig. 6. Precision-Recall curve for discomfort analysis. (a) Different methods for discomfort detection evaluated with Data-Clinic. (b) Different methods for discomfort detection evaluated with Data-Youtube.
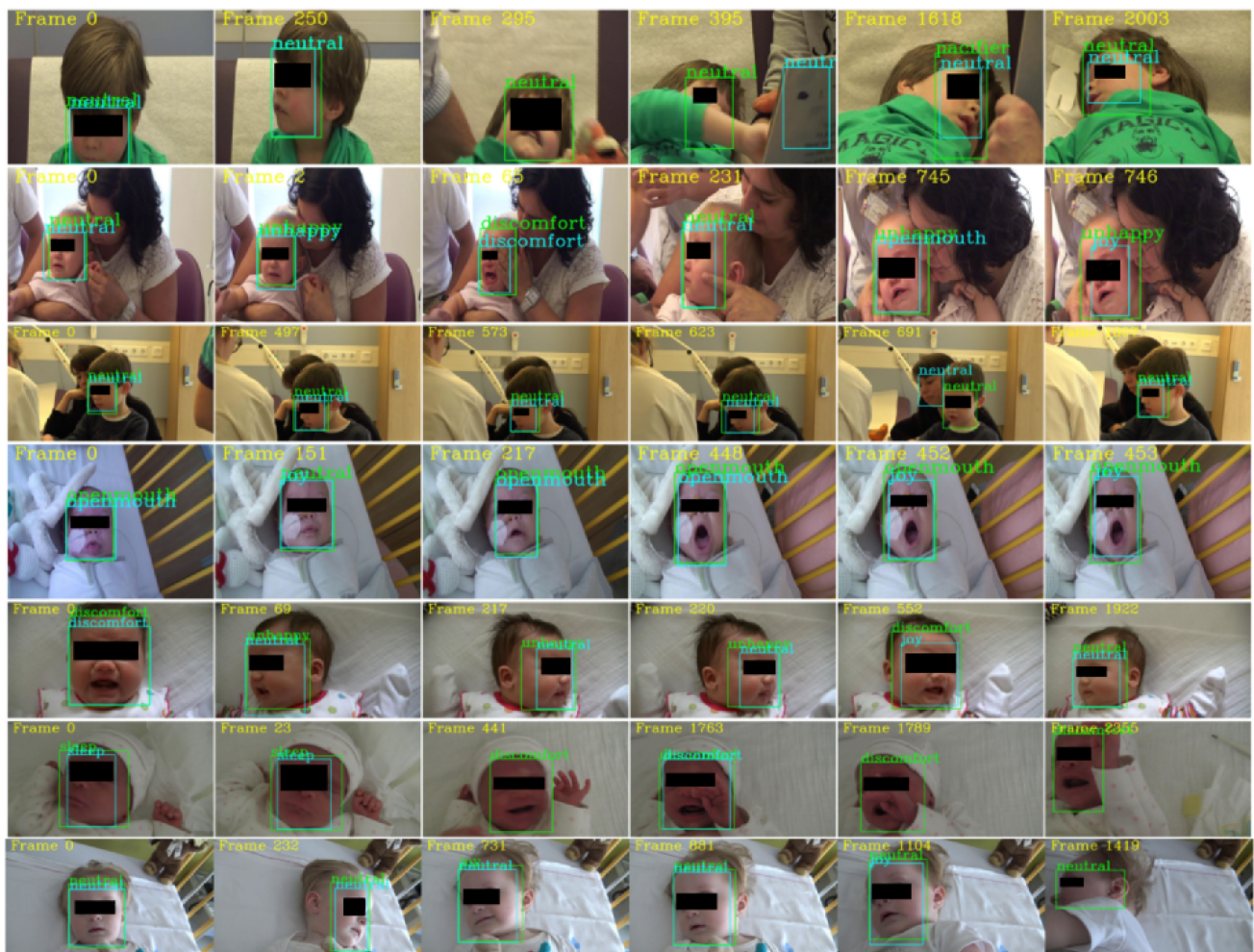

Fig. 7. Examples of expression detection on frames within sequences from the Data-Clinic dataset. Cyan boxes represent detections by Faster R-CNN, and green boxes represent detection by our R-CNN + HMM.

expressions which mainly appeared on faces of toddlers, such as suspect and disgust, the proposed framework trained with infant images will fail. Nevertheless, the detection of these expressions are outside the scope of this work, but can be studied in the future.

D. Computational Analysis

The computational cost for executing the proposed algorithm concentrates on Fast R-CNN and HMM. Compared with Faster R-CNN [6], our tracking has reduced the number of proposed ROIs from 300 to approximately 40 for each frame. For the HMM computation, the complexity for computing the update step is $O(n^2)$, where n is the number of proposed ROIs in the current frame. Since the number of proposed ROIs increases, the computation cost increases quadratically. In our experiment, the algorithm is executed on a GTX-1080ti GPU, which achieves a frame rate of 14 fps with 40 proposed ROIs per frame. When executing Faster R-CNN [6] on the same platform, we have obtained the performance of 9 fps. It can be concluded that the proposed method becomes feasible for a near real-time video-based infant monitoring system, for further prototyping and testing in a clinical environment. For large-scale deployment in hospitals or derivative systems intended for even consumer use at home, the current GPUs should become less expensive and require less power consumption, or embedded GPU systems should be used.

E. Discussion

Although increased stability 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 an HMM is combined with an unreliable classifier, the detection will be stuck at false positives. In the evaluational experiments with Data-Clinic, the HMM updates the detection score inclined to Pacifier rather than Neutral for a stable output, since the trained classifier behaves ambiguously between distinguishing a face with a pacifier and with objects presented near the mouth area as shown in Figure 7 (first row). Also, false positives occur when infants show a large pitch angle, so that the classifier will categorize the expression as sleep instead of a neutral face. However, these false positives can be further reduced by training the deep learning classifier and an HMM in an end-to-end fashion.

Despite the aforementioned limitations of the proposed framework introduced by HMM, the overall high performance of discomfort detection is encouraging for further experimental deployment of an infant monitoring system in a clinical environment at a somewhat larger scale. This larger-scale deployment is motivated and desired for three reasons. First, the monitoring system based on video analysis is not invasive and the installation of such a system is flexible and general to any hospital. In addition, the current execution speed allows the system to work in near real-time operation. With the adoption of an advanced GPU or applying parallel low-cost GPUs, the real-time performance can be already obtained. The real-time discomfort detection is already interesting for hospitals as an alerting machine for medical personnel. Second, because of the real-time operation,

a practical evaluation in several hospitals simultaneously helps in a fast large-scale validation of the proposed system and the large-scale clinical validation of the discomfort detection, so that the clinical benefits can become clear soon. Third, after this larger-scale evaluation, the experimental system can be modified and upgraded to improve clinical usage. An important near-future application is that the automated monitoring can assist a more-detailed diagnostics of diseases.

Actually, the proposed infant monitoring system is designed coupling to GERD diagnosis when the discomfort detection is combined with a pH-impedance measurement. However, this system can be extended to other disease diagnosis when a discomfort analysis is required, which adds significant value to the pediatric field. In the future, we will focus on training the expression classifier and the temporal analysis end-to-end for further improving the accuracy of the expression detection, and apply the proposed system for GERD diagnosis.

VII. Conclusion

In this paper, we have proposed a near real-time video-based infant monitoring system using Fast R-CNN combined with a dynamic model based on a Hidden Markov Model for increasing the stability of the decision making. Differentiating from the conventional methods that apply face detection and expression classification separately, we have trained a ConvNet detection framework oriented at infant expressions and states. The experimental results have shown an AP increase of 54.3% and 73.2% comparing with conventional methods for discomfort detection. When compared to Faster R-CNN, the proposed system outperforms up to 4.5%. It was shown that the proposed system becomes better when multi-class detection is employed up to seven categories, which facilitates clinical analysis of the infants and enables to combine these states with disease analysis like GERD. Moreover, the proposed system executes in near real time when implemented on a GPU. Since the R-CNN classifier is trained with single frames, false positives can occur for ambiguous expressions. However, with the proposed detection compensation method based on HMM, false positives can be reduced significantly. In the future, as more video sequences of infants become available, we will train our expression classifier and temporal analysis in an end-to-end fashion.

References

[1] M. Sullivan and M. Lewis, "Emotional expressions of young infants and children," *Infants Young Child.*, vol. 16, no. 2, pp. 120–142, Apr.–Jun. 2003.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA: IEEE Comput. Soc., 2014, pp. 580–587. [Online]. Available: https://doi.org/10.1109/CVPR.2014.81

[3] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4203–4212.

[4] J. Li *et al.*, "Facial expression recognition with faster R-CNN," *Procedia Comput. Sci.*, vol. 107, pp. 135–140, Dec. 2017.

[5] R. B. Girshick, "Fast R-CNN," *2015 IEEE Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[7] E. Fotiadou, S. Zinger, W. E. Tjon a Ten, S. Oetomo, and P. H. N. de With, "Video-based facial discomfort analysis for infants," in *Proc. SPIE - Int. Soc. Opt. Eng.*, vol. 9029, Jan. 2014, doi: 10.1117/12.2037661.

[8] C. Li, S. Zinger, W. E. Tjon a Ten, and P. H. N. de With, "Video-based discomfort detection for infants using a constrained local model," in *Proc. Int. Conf. Syst., Signals Image Proc.*, May 2016, pp. 1–4.

[9] Y. Sun *et al.*, "Video-based discomfort detection for infants," *Mach. Vis. Appl.*, Aug. 2018.

[10] J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, and P. J. Phillips, "Quantifying how lighting and focus affect face recognition performance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Workshops*, 2010, pp. 74–81.

[11] N. Witt, S. Coynor, C. Edwards, and H. Bradshaw, "A guide to pain assessment and management in the neonate," *Curr. Emerg. Hosp. Med. Rep.*, vol. 4, pp. 1–10, Mar. 2016.

[12] C. Hicks, C. Baeyer, P. Spafford, I. van Korlaar, and B. Goodenough, "The faces pain scale - revised: Toward a common metric in pediatric pain measurement," *Pain*, vol. 93, no. 2, pp. 173–183, 09 Aug. 2001.

[13] S. I. Merkel, T. Voepel-Lewis, J. R. Shayevitz, and S. Malviya, "The FLACC: A behavioral scale for scoring postoperative pain in young children," *Pediatr. Nurs.*, vol. 12, pp. 293–297, May/Jun. 1997.

[14] R. Zhi, G. Zamzmi, D. Goldgof, T. Ashmeade, and Y. Sun, "Automatic infants' pain assessment by dynamic facial representation: Effect of profile view, gestational age, gender, and race," *J. Clin. Med.*, vol. 7, no. 7, p. 173, Jul. 2018.

[15] K. Sikka *et al.*, "Automated assessment of children's postoperative pain using computer vision," *Pediatrics*, vol. 136, no. 1, pp. 124–131, Jul. 2015.

[16] S. Brahnam *et al.*, "Neonatal pain detection in videos using the icopevid dataset and an ensemble of descriptors extracted from gaussian of local descriptors," *Appl. Comput. Informat.*, Jun. 2020.

[17] G. Zamzmi, C. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in *Proc. 23rd Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 4148–4153.

[18] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[19] M. Tavakolian and A. Hadid, "Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 1952–1956.

[20] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 1957–1961.

[21] Z. Hammal, W. Chu, J. F. Cohn, C. Heike, and M. L. Speltz, "Automatic action unit detection in infants using convolutional neural network," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 216–221.

[22] A. Osmani, M. Hamidi, and A. Chibani, "Machine learning approach for infant cry interpretation," in *Proc. IEEE 29th Int. Conf. Tools Artif. Intell.*, 2017, pp. 182–186.

[23] S. Barajas-Montiel and C. A. Reyes-Garcia, "Identifying pain and hunger in infant cry with classifiers ensembles," Dec. 2005, pp. 770–775.

[24] R. V. Grunau and K. D. Craig, "Pain expression in neonates: Facial action and cry," *Pain*, vol. 28, no. 3, pp. 395–410, 1987.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv 1409.1556*

[26] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2009, pp. 248–255.

[29] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*.

[30] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, Jul. 2017, pp. 1387–1395.

[31] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, and Y. Sun, "A review of automated pain assessment in infants: Features, classification tasks, and databases," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 77–96, 2018. doi: 10.1109/RBME.2017.2777907.

[32] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Svm classification of neonatal facial images of pain," in *Fuzzy Logic and Applications,* I. Bloch, A. Petrosino, and A. G. B. Tettamanzi, Eds., Berlin, Heidelberg: Springer, 2006, pp. 121–128.

[33] D. Harrison *et al.*, "Too many crying babies: A systematic review of pain management of practices during immunizations on youtube," *BMC Pediatrics*, vol. 14, 2014, Art no. 134.

[34] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[35] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[36] C. Li, A. Pourtaherian, W. E. Tjon a Ten, and P. H. N. de With, "Infant monitoring system for real-time and remote discomfort detection," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2020, pp. 1–2.