E-BabyNet: Enhanced Action Recognition of Infant Reaching in Unconstrained Environments

Amel Dechemi, *Student Member, IEEE*, Konstantinos Karydis, *Member, IEEE*

***Abstract*— Machine vision and artificial intelligence hold promise across healthcare applications. In this paper, we focus on the emerging research direction of infant action recognition, and we specifically consider the task of reaching which is an important developmental milestone. We develop *E-babyNet*, a lightweight yet effective neural-network-based framework for infant action recognition that leverages the spatial and temporal correlation of bounding boxes of infants' hands and objects to reach for to determine the onset and offset of the reaching action. *E-babyNet* consists of two main layers based on two LSTM and a Bidirectional LSTM (BiLSTM) model, respectively. The first layer provides a pre-evaluation of the reaching action for each hand by providing onset and offset keyframes. Then, the biLSTM model merges the previous outputs to deliver an outcome of the reaching actions detection for each frame including the reaching hand. We evaluated our approach against four other lightweight structures using a dataset comprising 5,865 annotated images resulting in 16,337 bounding boxes from 375 distinctive infant reaching actions performed while sitting by different subjects in unconstrained (home/clinic) environments. Results illustrate the effectiveness of our approach and ability to provide reliable reaching action detection and offer onset and offset keyframes with a precision of one frame. Moreover, the biLSTM layer can handle the transition between reaching actions and help reduce false detections.**

***Index Terms*— Infant Action Recognition, Infant Reaching, Machine Vision, Artificial Intelligence**

I. Introduction

The ability to employ machine vision and artificial intelligence (AI) to identify, comprehend, and anticipate various human activities is critical to developing effective and interactive human-computer interfaces [1], [2] and healthcare systems [3]–[5]. Such an ability falls within the realm of developing vision-based human action recognition algorithms (e.g., [6]–[13]), often using different sensing modalities. Most existing works have considered the development of action recognition algorithms based on datasets that contain (young) adult motions (e.g., [14]–[17]) considering the large possible actions and application domains afforded by such populations.

This paper focuses on the emerging research thrust of *infant action recognition*, mainly toward pediatric applications. Some examples include identification of early signs of neuromotor disorders [18], [19] and assessment of the performance of behavioral and physical therapy through smart environments and assistive wearable robotic devices [20], [21]. Unlike datasets containing (young) adult activity, datasets containing infant motions are sparse due to stricter regulations protecting children's privacy as infants have no control over the release of the data, which has hindered the development of action recognition methods for this population.

There has been a keen effort in the research community to address the current shortage of infant activity datasets [21]–[26]. Such datasets contain information retrieved as RGB images, depth images, or 2D/3D skeletons through one or multiple cameras, and were collected following either passive or interactive paradigms to elicit infant motion. Motivated by an observed decrease in the accuracy of pose estimation approaches trained on adult data, Sciortino et al. [27] developed a dataset collected using publicly available videos of 104 children in natural environments and manually annotated 22 body keypoints. Another significant dataset is the Moving INfants In RGB-D (MINI-RGBD) based on the Skinned Multi-Infant Linear body model (SMIL) [23]. The authors used realistic shapes and textures to produce RGB and depth images for 2D and 3D joint positions by mapping real infants' movements to the SMIL model. Kokkoni et al. [20] designed an interactive infant learning environment where multiple Kinect cameras were used to capture various infant motions during both structured tasks and environment exploration. Recently, the InfAct (Infant Action) dataset was introduced in [25]. The dataset contains 200 video clips of infant activities and 400 images of infant postures. It also includes structured action and transition segmentation labels.

The focus of this work is on infant reaching action recognition, by specifically employing machine vision and AI as a means to achieve so. Reaching is a significant milestone since the ability of infants to explore and interact with their environment directly impacts their motor, social, perceptual, and cognitive development [28]–[31]. It is thus important to monitor and track the development of reaching skills. At the same time, new technology such as machine vision and AI can offer new ways to monitor children's development in fast, scalable, and accessible ways, sometimes even from in-home settings. This can be viewed as a form of telehealth/telemedicine, the significance of which was clearly demonstrated during the COVID-19 pandemic [32]. Hence, this work aims to develop a method for AI-based infant action

The authors are with the Dept. of Electrical and Computer Engineering, University of California, Riverside. (e-mail: {adech003, karydis}@ucr.edu). We gratefully acknowledge the support of NSF # CMMI-2133084. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

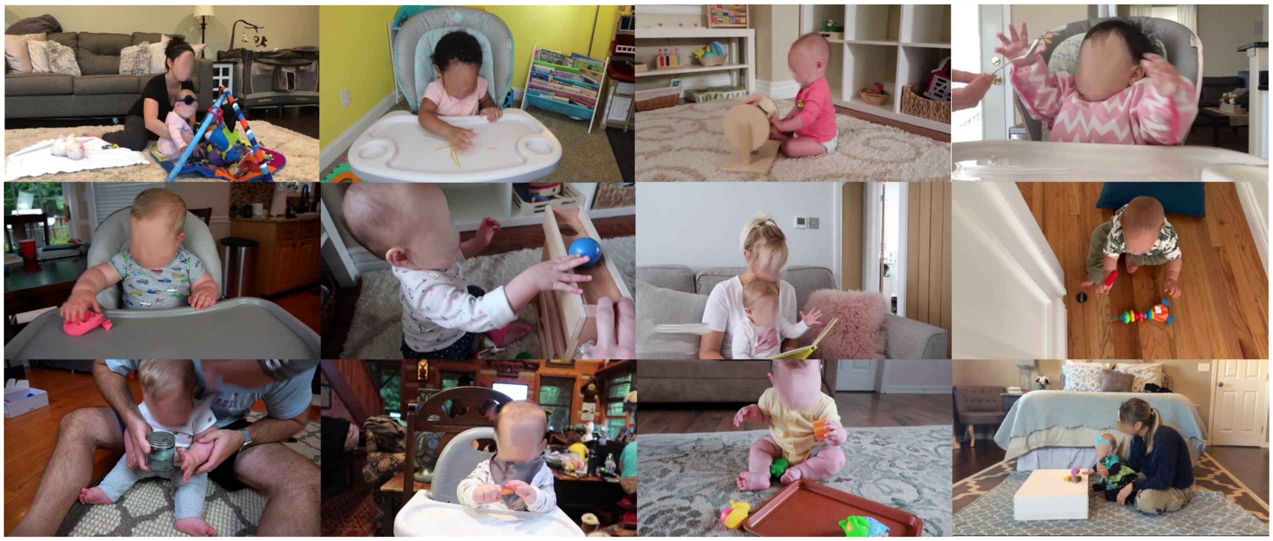

Fig. 1. Sample frames from the infant reaching dataset employed in this work.

recognition directly from videos taken at home (or in clinics) using regular video cameras. To do so, it is critical to employ datasets for training and validating the learning system.

Previous work [22] introduced a lightweight infant action recognition framework along with a reaching activity dataset that was obtained by curating videos published in the public domain via video-sharing platforms. Development of the dataset in that work was needed since experienced reachers tend to have more consistent reaching actions, thereby fundamentally differing from reaching actions of developing humans [33]. The learning structure proposed there was based on a Long-Short Term Memory (LSTM) model and provided the first and last frame corresponding to the reaching action and the reaching hand information. That work yielded important information which motivated the research in this present manuscript. Smaller memory-based structures were found to perform similarly with much larger CNN (ResNet-based [34]) structures. The latter exhibited strong performance during training and validation but resulted in increased false positive rates. Hence our previous work [22] demonstrated the preliminary feasibility and potential of employing small(er) structures. However, the consideration of a single layer alone revealed several limitations, including the inability to identify the active reaching hand and being sensitive to data sample types (i.e. left vs right hand reaching actions). In fact, missing hand/object detection can compromise the whole reaching action recognition.

To address those limitations, in this paper we propose a new two-layered learning structure able to assess separately right-handed and left-handed reaching actions as well as to handle both very short and bimanual reaches which were previously difficult to detect. This is facilitated by the first layer consisting of two independent LSTM modules similar to *BabyNet* [22] that operate in parallel and each is responsible for the left or the right hand. Then, we integrate a Bidirectional LSTM (biLSTM) structure in a second layer that fuses the outputs of the two parallel models and enables better identification of the transition between the no-reaching and the reaching phases. Further, we consider a larger dataset of annotated images and associated bounding boxes that feature examples from bimanual reaching actions, have diverse camera viewpoints, and include various occlusions and/or hand/object overlap. We test the efficacy of the new structure against several small memory-based baselines and assess our method's performance in a series of ablations studying the effect of key hyperparameters, out-of-sample generalization, and utility of the BiLSTM network as the second layer component.

II. Methods

A. Dataset of Infant Reaching Actions

This work employs an extended version of the dataset reported in [22]. The dataset contains videos from infants (up to 12 months of age) performing reaching actions in natural scenes (at home and/or in clinical settings)—sample cases are shown in Fig. 1. The videos were identified and collected via online publicly available video-sharing platforms. [1] From the viewpoint of visual learning, environmental conditions in the considered videos are relatively stable (e.g., good lighting and relatively clear backgrounds), but several factors introduce complexity in the dataset. Most critically, there are at cases occlusions, camera viewpoint variations between different subjects, accompanying adults' hands appearing in the scene, and multiple possible target objects to potentially reach for (see Fig. 1). Compared to [22] that considered 193 reaching actions from 21 subjects, this work considered an increased number of reaches (375) performed by 40 different infants which also included reaching actions from a wider variety of camera viewpoints as well as bimanual reaches.

Reaching actions were obtained via a manual annotation process. We followed the approach outlined in [22], whereby a reaching action is a sequence of video frames between the Reaching Onset (RN) and the Reaching Offset (RF). The onset was defined as the first frame in which movement of the infant's hand (left, right, or both) toward an object presented to the infant is initiated. Herein, we denote this frame as the

[1] We refer the interested reader to [22] for more details regarding the video collection process such as inclusion/exclusion criteria.

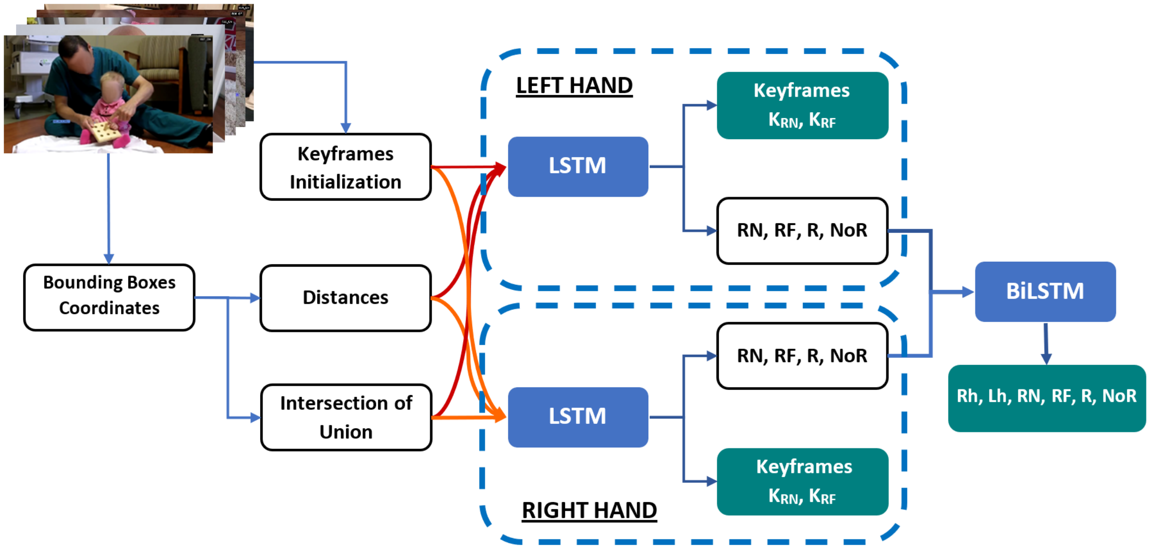

Fig. 2. Flowchart of our approach *E-BabyNet*. At initialization, both Reaching Onset and Offset Keyframes, $\boldsymbol{K_{RN}}$ and $\boldsymbol{K_{RF}}$, are set as the first frame of the sequence of frames $\boldsymbol{\mathcal{F}}$. For each pair $\boldsymbol{\mathcal{F}_j}$ and $\boldsymbol{\mathcal{F}_{j-1}}$, the bounding boxes of both left and right hands and the object are extracted. Next, the distances and the intersection of union are calculated for each frame and are provided as inputs to the first layer of the structure. Each subsystem assesses the reaching action for each hand by providing an output that consists of four probabilistic scores: Onset (RN), Offset (RF), Reach (R), and No Reach (NoR), and used to update the Onset and Offset Keyframes ($\boldsymbol{K_{RN}}$ and $\boldsymbol{K_{RF}}$). Then, the outputs of two parallel models in the first layer are merged in the second layer via a Bidirectional LSTM (biLSTM) network. This yields the final output including the four states RN, RF, R and NoR, along with the probabilities of Left (Lh) and Right (Rh) reaching hand.

Onset Keyframe (K_{RN}). The offset is the first frame in which any part of the infant's hand(s) make intentional contact with the object; we denote this frame as the *Offset Keyframe* (K_{RF}).

A total of N=375 reaches were considered; 158 for the left hand (LH) and 217 for the right hand (RH). Let $J^n \in \mathbb{N}, n = 1, \ldots, N$ denote the duration for each reaching action, i.e. the number of all frames, $\mathcal{F}_j, j = 1, \ldots, J^n$, between K_{RN} and K_{RF} keyframes. The duration, J^n, of LH and RH reaches are quite balanced, and follow a similar distribution. In detail, 59.24% of LH reaching actions are within a range of 3 to 15 frames (34.41% LH reaches last only between 3 to 7 frames). Longer reaches for the left-hand case are considerably fewer. LH reaches lasting between 31 to 48 frames and 49 to 63 frames comprise only 6.37% and 1.27% of the total number of left-hand reaches, respectively. RH reaching actions have about the same percentage of actions in the range of 3 to 15 frames as in the LH case. Actions with a length exceeding 30 frames represent up to 10% of the total number of right-hand reaches. Further, the number of left- and right-hand reaches is roughly the same for actions within the range of 3 to 7 frames, and for actions within the range of 8 to 15 frames.

Given all frames F_j for each reaching action, we developed bounding box annotations to track both the subject's hand(s) and any target objects (OBS). Annotations were done manually for each image throughout every reaching action.[2] A total of 5,865 images were annotated resulting in 16,337 bounding boxes. For each bounding box, the respective annotation provides the planar (x, y) coordinates for its top-left corner as well as its width and height (coordinates refer to the image plane, and all dimensions are in units of pixels). This information was employed to train, test, and assess the efficacy of our proposed action recognition approach as discussed next.

B. Development of E-BabyNet

Figure 2 illustrates the overall structure of our proposed approach for infant reaching action recognition, *E-BabyNet*. Let $\mathcal{B} = \{b_i^{\mathcal{F}}\}$ and $\mathcal{H} = \{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ denote the objects and the left and right hand detected in a sequence of frames $\mathcal{F}$, respectively. Each detection is associated with a bounding box with its top-left corner defined by x and y coordinates (in pixels); C_{b_i}, C_{h_L} and C_{h_R} denote the object, the left hand and the right hand bounding boxes, respectively. Let $D_{b_i} := \{H_{b_i}, W_{b_i}\}$ denote the bounding box dimensions for objects. Likewise, let $D_{h_L} := \{H_{h_L}, W_{h_L}\}$ and $D_{h_R} := \{H_{h_R}, W_{h_R}\}$ denote the bounding box dimensions for the left hand and the right hand, respectively.

At initialization, both keyframes, K_{RN} and K_{RF}, are set as the first frame of $\mathcal{F}$. The obtained bounding box detections are used to correlate spatiotemporal patterns between $\mathcal{H}$ and $\mathcal{O}$, and undergo two separate processes for the Reaching Onset phase (RN) and the Reaching Offset phase (RF). A qualitative example to demonstrate the process is shown in Fig. 3.

The Onset Phase: First, we compute the distance d_j^i between each hand $\{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ and the object $b_i^{\mathcal{F}}$ in the current frame $\mathcal{F}_j$. Next, we evaluate $d_j^i - d_{j-1}^i$ as follows: if $d_j^i - d_{j-1}^i < 0$, the onset is confirmed and $\mathcal{F}_{RN}$ is kept as the onset keyframe K_{RN}; however, if $d_j^i - d_{j-1}^i \geq 0$ and continues increasing for four consecutive frames, the current onset is invalidated and the onset keyframe K_{RN} is updated as $\mathcal{F}_j$. This is a practical way to help prevent false detections since the shortest reaching action lasts only three frames.

[2] Various learning-based object detection methods such as YOLO [35] or Mask R-CNN [36] are in principle applicable to help automate this process. Our method is currently agnostic to how these detections are made (which helps with generalization), and at the same time, the developed annotations can be employed to finetune/train such object detection networks.

The Offset Phase: To confirm the contact of the hand with the object, the Intersection of Union (IOU) is estimated between each hand $\{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ and the object $b_i^{\mathcal{F}}$. To do so, we extract the distance between the center of the LH and RH bounding boxes and the center of all possible target OBJ bounding boxes. The computed IOU is compared against a threshold value (a hyperparameter of our approach). This threshold can be estimated empirically via the data analysis process. The keyframe K_{RF} is updated as the current frame $\mathcal{F}_j$ as long as the IOU is less than the threshold and, thus no offset is confirmed. Otherwise, if contact is detected, the keyframe K_{RF} is definitively set as the current frame $\mathcal{F}_j$, and a new reaching action is initiated.

Structure of the E-BabyNet: With reference to Fig. 2, our visual learning structure employs the aforementioned information as input to a two-layer network. The first layer contains two sub-networks organized in parallel to each other. These sub-networks are LSTM models built directly upon our previous work *BabyNet* [22], and consider either the right or the left hand separately. They each aim at assimilating the correspondence between the corresponding hand's and the object's bounding boxes using the distances and IOU. Each sub-network assesses the reaching action for the corresponding hand via an output comprising four probabilistic scores: Onset (RN), Offset (RF), Reach (R), and No Reach (NoR). The Onset (RN) assesses the probability that the hand is about to initiate the action, while the Offset (RF) represents the probability that the reaching action is completed. All frames that are between the Onset (RN) and the Offset (RF) phases are marked as Reach (R), whereas the frames before the Onset and after the Offset are labeled as No Reach (NoR).

The outputs of the two parallel sub-networks in the first layer are fused into the second layer of our structure via a Bidirectional LSTM (biLSTM). The latter processes the input in a forward and a backward direction. Thus, it considers past and future information critical when transitioning from a Reach (R) to a No Reach (NoR). The final output is six-dimensional, and includes the four probabilistic scores RN, RF, R, and NoR (as defined previously) along with two additional probabilistic scores of Left (Lh) and Right (Rh) hand. These two scores highlight the probability that the left or right hand, respectively, is the reaching one. When both values are increased at the same time, this indicates a bimanual reaching action. To illustrate this output structure, consider the following exemplary case. A correctly detected reaching action with the right hand will yield a high score for Rh throughout the reaching action's duration (i.e. all frames). The RN score will spike at and around true onset (and be low in other frames), while the RF score will spike at and around true offset (and be low in other frames). Finally, the score for R will be high in all frames between RN and RF, and the NoR score will be high in all frames before RN and after RF.

C. Implementation and Experiments

To assess the performance of our approach, we considered an implementation and an evaluation phase. During the former, we considered a split of the dataset as 70% training, 15% validation, and 15% testing using 266 reaches out of the total of 375. We performed a comparison against various baselines and conducted an ablation study to determine the effect of key hyperparameters on the performance of *E-BabyNet*. For the latter phase, we used the remaining 109 reaches, consisting of 58.72% reaching actions lasting between 3 to 15 frames and 35.78% lasting between 16 to 30 frames. Besides evaluating the complete *E-BabyNet* network structure, we also performed an additional ablation study where the first layer models for the left hand and the right hand were evaluated independently on this unseen part of the dataset, to assess the significance and necessity of the second layer (BiLSTM). All experiments were performed on a workstation featuring an Intel Core i7 Processor (16 x 2.30GHz), 16 GB DDR4 RAM, and an NVIDIA GeForce RTX 3050 GPU. All networks were trained with learning rate 10e-4 using Adam optimizer and cross-entropy loss with 30 hidden layers.

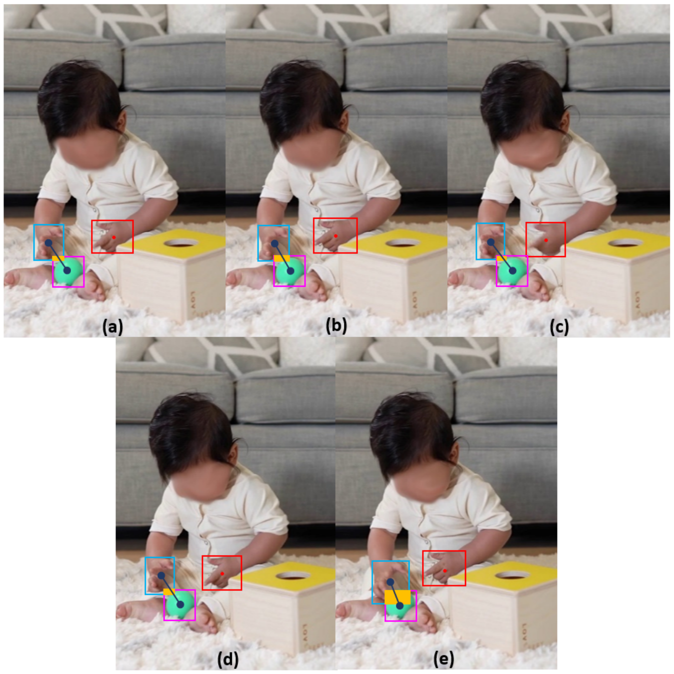

Fig. 3. Qualitative illustration of our approach. (a) Onset: the movement of the right hand (cyan) is initiated to reach the object (magenta). During the (b)-(d) reach and (e) offset: the distance between the right hand and the object (dark blue) decreases as the IOU (orange) increases until its value stabilizes (which indicates that a reaching action has been successfully achieved). Note that during this specific example, the distance between the left hand (red) and the object does not decrease in a significant manner and hence the respective IOU remains null which confirms that no reaching action is performed by the left hand.

Based on [22] that demonstrated the efficacy of comparatively smaller (and memory-based) networks,[3] the baselines in this work include only lightweight structures. These are:

- Multi-Layer Perceptron (MLP),
- Gated Recurrent Unit (GRU),
- Bidirectional LSTM (BiLSTM), and
- *BabyNet* (based on LSTM).

[3] In earlier work we evaluated several large CNN methods (specifically, ResNet-based [34] architectures combined with different data augmentation techniques), and it was found [22, Table II] that these larger structures demonstrated solid performance during training and validation but were providing results with increased false positives rates.

TABLE I
PERFORMANCE OF THE *E-BabyNet* AND COMPARATIVE RESULTS AGAINST BASELINES.

Model	Parameters	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
MLP	183	64.7	59.1	60.3	0.66	0.61	0.80
GRU	3153	72.6	72.3	85.6	0.68	0.70	0.82
BiLSTM	8103	76.3	77.7	85.5	0.71	0.69	0.81
BabyNet	3960	71.8	72.1	89.5	0.70	0.69	0.82
E-BabyNet	8249	95.4	97	95.5	0.96	0.97	0.97

TABLE II
HYPERPARAMETERS EFFECT ON THE PERFORMANCE OF THE *E-BabyNet*.

Batch							
Batch	Hidden Layer	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
16	30	89.4	83.7	94.7	0.82	0.82	0.87
32	30	73.7	74.9	95.2	0.88	0.95	0.96
64	30	96.2	97.5	97.3	0.98	0.97	0.98
full	30	95.4	97	95.5	0.96	0.97	0.97
Hidden Layer							
Batch	Hidden Layer	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
full	15	61.5	61.0	57.6	0.82	0.82	0.87
full	20	72.6	72.3	85.6	0.68	0.70	0.82
full	30	95.4	97	95.5	0.96	0.97	0.97
full	100	99.9	100	100	1.0	1.0	1.0

The aforementioned baselines employ two inputs, that is, the distances and the IOU of the detected hands' and objects' bounding boxes, and output the probabilistic scores (RN, RF, R, NoR) and the final keyframes K_{RN} and K_{RF}. *E-BabyNet* employs the same input format and outputs probabilistic scores Rh, Lh, RN, RF, R, NoR (see Fig. 2).

III. RESULTS

Obtained experimental results during the two phases are presented in Table I (*E-BabyNet* performance and comparison against baselines), Table II (ablation study regarding hyperparameters), and Table III (evaluation in unseen part of the dataset and ablation study regarding the second layer). In the implementation phase, we provide classification accuracy for training, validation, and testing for each tested structure. In both phases, we evaluated the precision, recall and area under the receiver operating characteristic (ROC) curve (AUC) to assess the tested structure's performance.

A. Implementation Phase

Comparison Against Baselines: With reference to Table I, it can be readily observed that the GRU, *BabyNet*, and BiLSTM baseline structures yielded similar performance. Their average testing accuracy exceeds 85%, but the recall score is at 0.70, which essentially indicates a higher number of false positives. The MLP demonstrated the worst performance, especially in terms of average accuracy and recall scores. The worse quantitative performance can also be associated with observed qualitative results. For all the correct reaching actions, the MLP had a difference within a range of five to eight frames in its keyframes. Moreover, it was unsuccessful at detecting comparatively shorter reaching actions (lasting between 5 to 8 frames) compared to the three other baseline structures.

In contrast, *E-BabyNet* yielded the best performance, with an average testing accuracy of 95.5% and high precision (0.96) and recall (0.97) scores. The AUC score confirms these findings as *E-BabyNet* has a score of 0.97. *E-BabyNet* had a delay in a range of one to four frames but was able to yield fewer false positives compared to all other methods.

Effect of Hyperparameters in E-BabyNet: The key hyperparameters considered herein include the batch size and the number of hidden layers. With reference to Table II, it can be seen that as the batch size varies, *E-BabyNet* can maintain its performance albeit with somewhat lower accuracy and smallest recall of 0.82 for the smallest batch size. The full and 64 batch sizes performed best with high recall and precision scores; however, the full batch size needed over eight times less time to train, providing a better trade-off between performance and time execution. With respect to the number of hidden layers, it can be observed that the number of hidden layers has a much more immediate and variable effect on the network's performance. A smaller number of hidden layers (20 or less) leads to a noticeable drop in performance, while the structure will overfit with a hidden layer of 100 after only 12 epochs. Based on these findings, we inferred that the structure with 30 hidden layers and full batch size is the most appropriate for our application.

B. Evaluation Phase

With reference to Table III, *E-BabyNet* can maintain its performance even in the unseen part of the dataset, with precision and recall scores of 0.73 and 0.82, respectively. The reported keyframes had an average precision of three frames. Assessing how the method would perform if the second layer (the BiLSTM) was not employed yields variable results. The left hand model has worse performance compared to the right hand model (and the complete network). However, the right hand model alone has an overall better performance than the complete network. These findings suggest that when the outputs of the two parallel models for the left and right hand

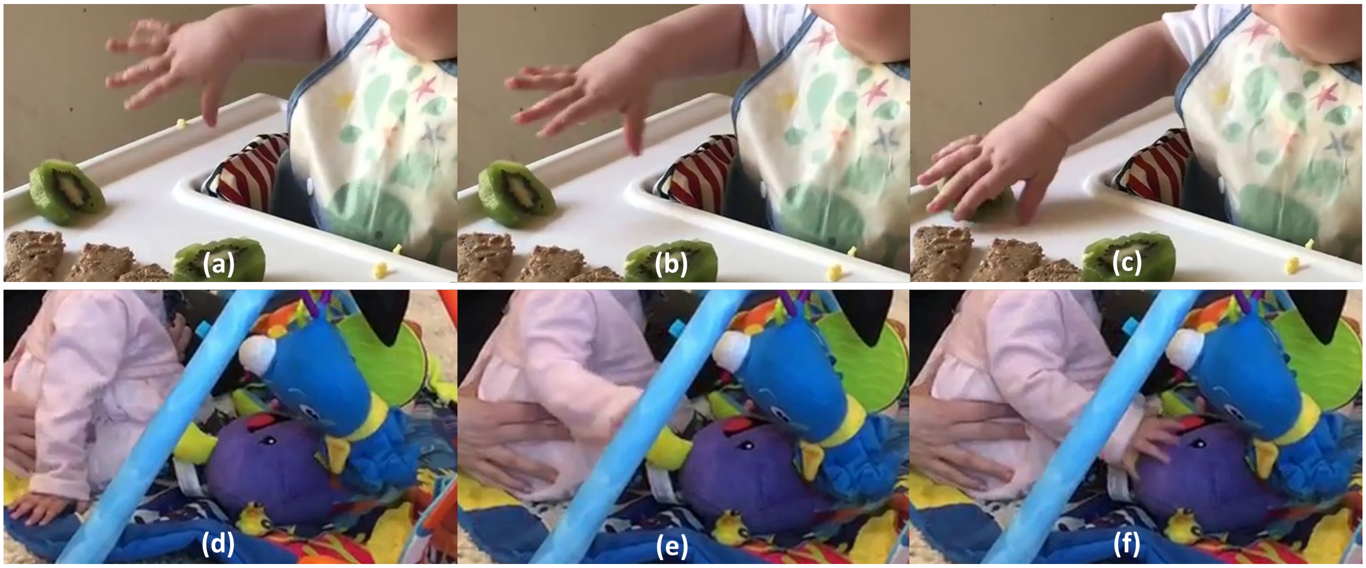

Fig. 4. Examples of occlusion of the infant's hand during reaching actions. The top panels (a, b, and c) show the object obstructed by the infant's hand. The bottom panels (d, e, and f) show a hand obstructed by an object.

are pooled together via the BiLSTM in the second layer, the network can become more robust. We observed that 72% and 79% of reaching actions lasting between 3 to 7 and 8 to 15 frames were correctly recognized, respectively. In the case of challenging or confusing cases, the *E-BabyNet* is more prone to predict a false negative than a false positive. Lastly, out of the six reaching actions with frames number greater than 31, four were correctly detected with a precision of four frames.

IV. Discussion

The obtained results collectively suggest that the developed *E-BabyNet* can demonstrate solid performance. The average testing accuracy of our method was 95.5%, with a precision of 0.96 and a recall of 0.97. Additional experimentation by evaluating our method using a never-seen-before part of the dataset yielded solid results with precision and recall at 0.73 and 0.82, respectively. These findings show that it is possible to perform robust infant reaching action recognition with comparatively small learning-based structures.

Three main features can be associated with this solid performance. First, the integration of memory is important as it helps capture the temporal correlation between the onset and offset points of a reaching task. This assessment can be made best when compared to our earlier results in [22, Table II] where large, CNN-based structures had lower performance compared to smaller, memory-based structures. Second, among smaller memory-based networks, the ability to differentiate between left-hand and right-hand reaches and assess them separately can help increase performance (i.e. average testing accuracy, precision, and recall). This is best viewed by comparing the first four baseline cases in Table I (all of which do not explicitly differentiate the left- and right-hand reaches) with the fifth line of the same table and with Table I. Results when both hands are considered explicitly are in all cases better. Third, and related to the previous feature, is the addition of the BiLSTM in the second layer to fuse the information from the first layer that addresses the left and right hand separately. This is critical because it reduces sensitivity to data sample types (i.e. left vs right hand reaching actions) as shown in Table III, while it also considers past and future information critical when transitioning from a Reach (R) to a No Reach (NoR). Owing to the BiLSTM integrated into the second layer of *E-BabyNet*, discontinued reaching actions were predicted regardless of the number of frames with partial information, e.g., the hands could be obstructed by the infant's body or another object in the scene during the reaching actions (Fig. 4). Predictions of keyframes were also highly accurate, with some delays observed only in reaching actions that also included a final grasp of the object.

TABLE III
Performance of the *E-BabyNet* During the Evaluation Phase.

Model	Precision	Recall	AUC
Left	0.69	0.89	0.86
Right	0.75	0.91	0.92
E-BabyNet	0.73	0.82	0.85

Besides improved performance, *E-BabyNet* has additional core benefits compared to baselines. While the baseline models underperformed in the case of short reaching actions, the *E-BabyNet* can recognize short reaching actions as well as long ones. It can also recognize bimanual reaching actions. Results also suggest that the baseline structures provided low scores for precision which can lead to high false positives. In contrast, *E-BabyNet* can provide reliable action recognition, and when uncertainty is too high, it will render a false negative detection. This last point is critical in the context of our application that seeks to integrate visual-learning-based reaching action recognition into a new soft wearable robotic device [37]–[41]. Our soft wearable robotic device seeks to offer perceptual-based assistive feedback to infants with upper extremity mobility impairments, and for safety purposes, it is important in the case of failing to correctly recognize a reaching action to not force an infant's arm to move against their will. (A false negative will not lead to device actuation.)

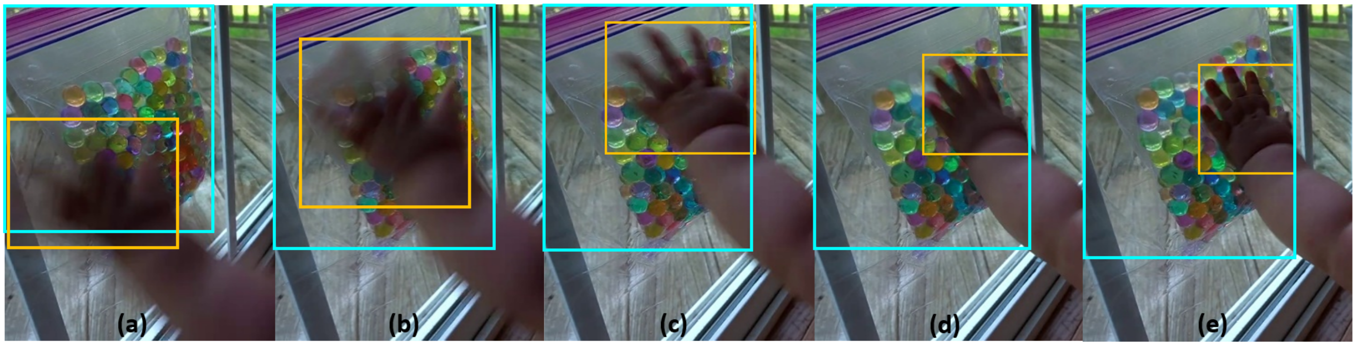

Fig. 5. A reaching action with overlapping hand (orange) and object (cyan). (a) Onset phase, (b)-(d) during reaching action, and (e) Offset phase.

Despite these positive findings, the evaluation phase highlighted lingering challenges that can be addressed to further improve our method's outcomes. Some of these challenges pertain to the complexity of the visual learning task considered herein. This is because of various types of possible occlusions and hand/object overlap (Fig. 4 and Fig. 5), camera viewpoint variations among different subjects, as well as accompanying adults' hands and multiple possible target objects appearing in images (Fig. 1). The front-view reaching actions were the most challenging for the *E-BabyNet* as the bounding boxes of the hand and the object may have significant overlap (Fig. 5). This effect was noticeable on the low score of the left model since left reaching actions were largely captured from the front (or rear) camera viewpoints. In this case, the IOU is set at a high value during the onset phase and thus infers either a shorter reach or a no reach. Further, reaching actions of developing humans can be less straight, less smooth, and more diverse [33], and this can lead to cases where the distance between the hand and the object increases significantly throughout the reaching action. This impacts the detection of reaching if it occurs for more than three frames. Figure 6 illustrates two such cases. We also observed that several grasping actions were reported by the network as featuring a reduced number of frames compared to the correct one when the infants' hand contacted the target object first before grasping it. This can be associated with the fact that the employed dataset includes training offsets with both touching and grasping, which can lead to a few frames difference in the offset although the detection remains correct.

While further improving out-of-sample outcomes is possible (e.g., by implementing more robust metrics for reaching offset determination that employ additional information such as 2D skeleton data), the current findings show that it is possible to perform robust infant reaching action recognition with comparatively small learning-based structures. We believe that this work can help inspire more works in this emerging area of infant reaching action recognition via visual learning and serve as a basis for future works to build on top of it.

V. Conclusion

In this work we proposed *E-BabyNet*, a learning-based structure that leverages machine vision and artificial intelligence to recognize infant reaching actions in unstructured environments. Results obtained through different types of testing and evaluation demonstrated the efficacy of *E-BabyNet* as compared to other lightweight neural-network-based structures. This work also enables interesting future directions of research. First, the rich infant motion variability during development could be better harnessed by further extending the reaching action dataset. To address challenging camera views, fusion with the infant's 2D skeleton data (at a minimum their arms, if not full-body) may improve the result of the first layer of the *E-BabyNet* structure, and thus increase the overall reaching action recognition efficiency. Furthermore, an online learning scheme could expand the scope of the structure and its capability to detect complex reaching actions, while a prediction model might aid in reducing the effect of hands' and objects' occlusions during reaching actions. It is also possible to integrate automated hand detection processes, for instance via learning-based visual detectors. In addition, the precision of the keyframe detection could be enhanced by an explicit definition of grasping and touching objects.

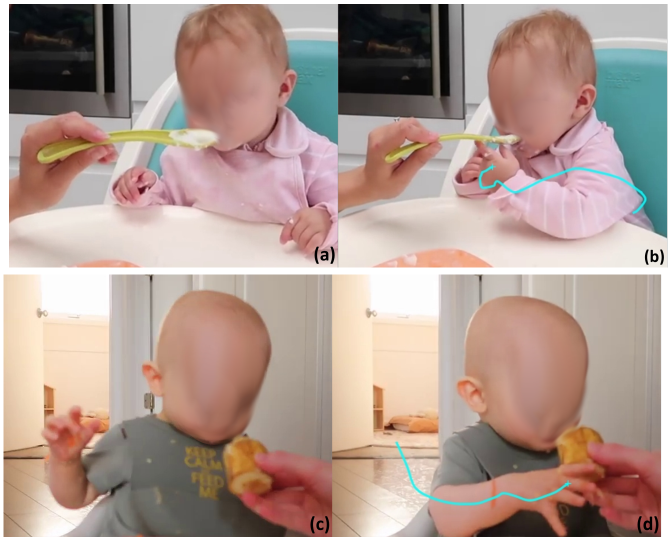

Fig. 6. Examples of reaching action trajectories of a (top) 6 and (bottom) 7 months infant. Panels (a) and (c) correspond to onset phases, while panels (b) and (d) show frames at the offset phase along with the complete trajectories followed by the hand during the reaching action.

VI. Acknowledgments

The authors thank Dr. Elena Kokkoni for her feedback and insights during the paper's development, as well as Vikarn Bhakri, Arjun Modi, Julya Mestas, Pamodya Peiris, and Dannya Enriquez Barrundia for their assistance at the initial stages of developing this work.

VII. REFERENCES

[1] D. Webster and O. Celik, "Systematic review of kinect applications in elderly care and stroke rehabilitation," *Journal of Neuroengineering and Rehabilitation*, vol. 11, no. 1, pp. 1–24, 2014.

[2] J. Choi, Y.-i. Cho, T. Han, and H. S. Yang, "A view-based real-time human action recognition system as an interface for human computer interaction," *Virtual Systems and Multimedia*, vol. 1, pp. 112–120, 2007.

[3] A. Ghali, A. S. Cunningham, and T. P. Pridmore, "Object and event recognition for stroke rehabilitation," in *SPIE Visual Communications and Image Processing*, vol. 5150, 2003, pp. 980–989.

[4] N. Alshurafa, W. Xu, J. J. Liu, M.-C. Huang, B. Mortazavi, C. K. Roberts, and M. Sarrafzadeh, "Designing a robust activity recognition framework for health and exergaming using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1636–1646, 2014.

[5] B. Debnath, M. O'brien, M. Yamaguchi, and A. Behera, "A review of computer vision-based approaches for physical rehabilitation and assessment," *Multimedia Systems*, vol. 28, no. 1, pp. 209–239, 2022.

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[7] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision*, 2018, pp. 720–736.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *IEEE/CVF International Conference on Computer Vision*, 2011, pp. 2556–2563.

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[10] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *ArXiv*, vol. 1212.0402, 2012.

[11] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5842–5850.

[12] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. S. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *European Conference on Computer Vision*, 2020.

[13] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200–3225, 2023.

[14] J. Collins, J. Warren, M. Ma, R. Proffitt, and M. Skubic, "Stroke patient daily activity observation system," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2017, pp. 844–848.

[15] A. Jaume-i Capó and A. Samčović, "Vision-based interaction as an input of serious game for motor rehabilitation," in *IEEE Telecommunications Forum Telfor (TELFOR)*, 2014, pp. 854–857.

[16] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[17] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[18] D. Sakkos, K. D. Mccay, C. Marcroft, N. D. Embleton, S. Chattopadhyay, and E. S. Ho, "Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy," *IEEE Access*, vol. 9, pp. 94 281–94 292, 2021.

[19] L. Adde, A. Brown, C. Van Den Broeck, K. DeCoen, B. H. Eriksen, T. Fjørtoft, D. Groos, E. A. F. Ihlen, S. Osland, A. Pascal *et al.*, "In-motion-app for remote general movement assessment: A multi-site observational study," *BMJ open*, vol. 11, no. 3, 2021.

[20] E. Kokkoni, E. Mavroudi, A. Zehfroosh, J. C. Galloway, R. Vidal, J. Heinz, and H. G. Tanner, "Gearing smart environments for pediatric motor rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 17, no. 1, 2020.

[21] C. Pacheco, E. Mavroudi, E. Kokkoni, H. G. Tanner, and R. Vidal, "A detection-based approach to multiview action classification in infants," in *International Conference on Pattern Recognition*, 2021, pp. 6112–6119.

[22] A. Dechemi, V. Bhakri, I. Sahin, A. Modi, J. Mestas, P. Peiris, D. E. Barrundia, E. Kokkoni, and K. Karydis, "Babynet: A lightweight network for infant reaching action recognition in unconstrained environments to support future pediatric rehabilitation applications," in *IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 461–467.

[23] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. Sebastian Schroeder, "Computer vision for medical infant motion analysis: State of the art and RGB-D data set," in *European Conference on Computer Vision Workshops*, 2018, pp. 32–49.

[24] X. Cao, X. Li, L. Ma, Y. Huang, X. Feng, Z. Chen, H. Zeng, and J. Cao, "Aggpose: Deep aggregation vision transformer for infant pose estimation," in *International Joint Conference on Artificial Intelligence*, 2022, pp. 5045—5051.

[25] X. Huang, L. Luan, E. Hatamimajoumerd, M. Wan, P. D. Kakhaki, R. Obeid, and S. Ostadabbas, "Posture-based infant action recognition in the wild with very limited data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 4911–4920.

[26] E. Hatamimajoumerd, P. D. Kakhaki, X. Huang, L. Luan, S. Amraee, and S. Ostadabbas, "Challenges in video-based infant action recognition: A critical examination of the state of the art," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 21–30.

[27] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, "On the estimation of children's poses," in *International Conference on Image Analysis and Processing*, vol. 10485, 2017, pp. 410–421.

[28] K. E. Adolph and J. M. Franchak, "The development of motor behavior," *WIREs Cognitive Science*, vol. 8, no. 1-2, 2016.

[29] J. L. Williams and D. Corbetta, "Assessing the impact of movement consequences on the development of early reaching in infancy," *Frontiers in Psychology*, vol. 7, p. 587, 2016.

[30] M. A. Lobo and J. C. Galloway, "The onset of reaching significantly impacts how infants explore both objects and their bodies," *Infant Behavior and Development*, vol. 36, no. 1, pp. 14–24, 2013.

[31] M. Lobo, J. Galloway, and J. C. Heathcock, "Characterization and intervention for upper extremity exploration & reaching behaviors in infancy," *Journal of Hand Therapy*, vol. 28, no. 2, pp. 114–125, 2015.

[32] V. Sabesan, F. Ogunfuwa, J. Grunhut, S. Sommerville, C. Fomunung, J. Elkhechen, C. Fernandez, A. Lavin, and G. R. Jackson, "Telemedicine in orthopaedics during the covid-19 pandemic: a comparative landscape," *International Orthopaedics*, pp. 1–7, 2024.

[33] N. E. Berthier and R. Keen, "Development of reaching in infancy," *Experimental Brain Research*, vol. 169, pp. 507–518, 2006.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[37] E. Kokkoni, Z. Liu, and K. Karydis, "Development of a soft robotic wearable device to assist infant reaching," *Journal of Engineering and Science in Medical Diagnostics and Therapy*, vol. 3, pp. 021 109–1, 2020.

[38] I. Sahin, J. Dube, C. Mucchiani, K. Karydis, and E. Kokkoni, "A bidirectional fabric-based pneumatic actuator for the infant shoulder: Design and comparative kinematic analysis," in *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 371–376.

[39] C. Mucchiani, Z. Liu, I. Sahin, J. Dube, L. Vu, E. Kokkoni, and K. Karydis, "Closed-loop position control of a pediatric soft robotic wearable device for upper extremity assistance," in *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 1514–1519.

[40] I. Sahin, M. Ayazi, C. Mucchiani, J. Dube, K. Karydis, and E. Kokkoni, "A fabric-based pneumatic actuator for the infant elbow: Design and comparative kinematic analysis," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2023, pp. 1–6.

[41] C. Mucchiani, Z. Liu, I. Sahin, E. Kokkoni, and K. Karydis, "Robust generalized proportional integral control for trajectory tracking of soft actuators in a pediatric wearable assistive device," in *In IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2023, pp. 559–566.