UNIVERSITY OF CALIFORNIA
RIVERSIDE

Infant Reaching Action Recognition in Unconstrained Environments

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Vikarn Bhakri

June 2021

Thesis Committee:
Dr. Konstantinos Karydis, Chairperson
Dr. Elena Kokkoni
Dr. M. Salman Asif

The Thesis of Vikarn Bhakri is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Infant Reaching Action Recognition in Unconstrained Environments

by

Vikarn Bhakri

Master of Science, Graduate Program in Electrical Engineering
University of California, Riverside, June 2021
Dr. Konstantinos Karydis, Chairperson

Action recognition can play a key role in the automation of devices, such as those developed for physical rehabilitation. The majority of current work focuses on adult action recognition, and only a limited body of action recognition work is geared toward pediatric populations, such as infants. This work introduces a new lightweight neural network structure, BabyNet, that recognizes reaching motion of infants from off-body stationary cameras. This approach makes use of the spatial and temporal connection between bounding boxes around an infant's hands and object of interest in order to recognize a reaching action toward that object. BabyNet is trained and tested on a new dataset that showcases reaches performed in a sitting position by different infants in unconstrained environments, such as the families' homes. To evaluate the efficacy of the proposed approach, an ablation study is conducted where BabyBet is compared against several other learning-based architectures. Results show that BabyNet performs satisfactorily in terms of average testing accuracy by exceeding that of competing networks. Due to its small size, it can serve as a lightweight architecture for video-based infant reaching action recognition.

Contents

1 INTRODUCTION 1

1.1 Motivation 1
1.2 Challenges 2
1.3 Contributions 3

2 RELATED WORKS 4

2.1 Overview of Human Action Recognition Approaches 4
2.2 Use of Human Action Recognition for Rehabilitation Purposes 6

3 METHODOLGY 8

3.1 Dataset 8
3.2 Video Annotations 9
3.3 Bounding Box Annotations 13
3.4 Infant Reaching Action Algorithm 14

4 EXPERIMENTS 17

4.1 ResNet 17
4.2 ResNet+LSTM 18
4.3 3D Networks 18
4.4 LSTM with Optflow (O-LSTM) 19
4.5 Models Trained Using Bounding Box Coordinates 20

5 RESULTS 21

5.1 Effect of Data Augmentation 21
5.2 Effect of Integrating LSTM Block 22
5.3 Performance of the 3D Networks 23
5.4 Comparison of The BabyNet and The MLP Network 23
5.5 Performance of the O-LSTM Network 24

CONCLUSION 25

REFERENCES 26

List of Figures

Figure 1. Examples of Reaching Onset (left) and Offset (Right) Frames 10

Figure 2. Distribution of Reaching Actions 11

Figure 3: Sample Frames from the Dataset 14

Figure 4: Approach Used in the Infant Reaching Action Algorithm 15

Figure 5: The Images on the Top Depict the RGB Images and the Images on the Bottom Show the Equivalent Optical Flow Images 22

List of Tables

Table 1: Categories and Subcategories Annotated to Describe a Reaching Action .. **Error! Bookmark not defined.**2

Table 2: Subject Information and Corresponding No.of Reaches and Objects Annotated .....**Error! Bookmark not defined.**2

Table 3: Evaluation of the Models Trained on the Infant Reaching Dataset..... 21

Table 4: Evaluation of the 3D Networks Trained on the Infant Reaching Dataset 23

Chapter 1

Introduction

1.1 Motivation

According to the Centers for Disease Control and Prevention, traumatic brain injuries (e.g., cerebral palsy) are the leading cause of disabilities among infants in the U.S (Centers for Disease Control and Prevention, 2014). The highest incidence rates of traumatic brain injuries are found in children between birth and 4 years of age (Taylor et al., 2017). Such injuries may have negative effects on children's motor development as well as other developing systems, such as perception and cognition (Campos et al., 2000). Studies have shown that early intervention provided from birth to 3 years can help improve the motor function of infants with disabilities (Nectac, 2011). Therefore, it is of tremendous importance to develop technology that aids in diagnosing, assessing, and rehabilitating infants with brain injuries and/or motor disabilities early in life (Arnold et al., 2020).

Recently, computer vision research efforts have been geared toward the recognition of human motor actions that emerge early in life, such as infants' spontaneous and/or fidgeting movements, kicking, crawling, etc. (Chambers et al., 2020; Emeli et al., 2020; Pacheco et al., 2021). Being able to successfully automate the processes of detecting, recognizing, and classifying actions performed by infants from visual data such as images and videos can be useful in several pediatric applications. Some

examples include monitoring for infants' safety (Goto et al., 2013; Nie et al., 2018; Suzuki et al., 2012), identifying markers for diagnosis of neuromotor disorders (Chambers et al., 2020; Hashemi et al., 2014; McCay et al., 2020; Stahl et al., 2012; Westeyn et al., 2012), and lastly creating smart environments, automated assistive devices and autonomous robots for pediatric rehabilitation (Efthymiou et al., 2018; Hadfield et al., 2019; Kokkoni et al., 2020; Marinoiu et al., 2018; Tsiami et al., 2018).

1.2 Challenges

Developing algorithms that can accurately and reliably recognize actions performed by infants is not a straightforward process and has several challenges when compared to action recognition in adults. First, in the case of young children and infants, there is a lot of inherent variability in the movements and actions performed by the different age groups as a natural result of learning and growth (Fetters, 2010). For example, the reaching kinematic profiles of infants (the motor action of interest to this work) are more variable (Konczak & Dichgans, 1997), and it can take years for infants to achieve smooth and straight reaching trajectories similar to those seen in adults (Berthier & Keen, 2006; Schneiberg et al., 2002). Another difference encountered in comparison to adults is that infants can attain poses not common in adults (Sciortino et al., 2017). Therefore, using skeletal models or learning-based pose estimation methods typically used in adult action recognition may not be the best approach. The latter was also suggested in previous work in which models based on adult action recognition under-performed when tested on datasets from children (Efthymiou et al., 2018; Suzuki et al., 2012).

The second challenge is related to the lack of human activity datasets that contain sufficient and variable infant body poses and actions. Typically, any state of the art action recognition method requires training on an extensive dataset (Carreira & Zisserman, 2017; Goyal et al., 2017; Gupta et al., 2020; Kuehne et al., 2011; Soomro et al., 2012) and various datasets of adult activities are available as these can be used in a wide range of applications (Hesse et al., 2019; Pacheco et al., 2021). However, this is not the case for infant datasets. This can be due to the increased access restrictions placed for the protection of minors and/or the lower number of applications that are available for that age range, compared to adults. Therefore, this work aims to provide new avenues to overcome the aforementioned challenges.

1.3 Contributions

This thesis contributes to the field of infant action recognition in the following ways:

- A new lightweight neural network structure is proposed for infant reaching action recognition. The network is built upon an LSTM module to model the various stages of the reaching action through a spatio-temporal interpretation.
- A different network that uses optical flow images is also developed and explored in the context of infant action recognition.
- An ablation study is conducted with various structures evaluated on a dataset developed in parallel collaborations to compare the performance of the proposed network.

Chapter 2

Related Works

2.1 Overview of Human Action Recognition Approaches

There has been a steady shift from the use of traditional methods such as Support Vector Machines or Hidden Markov Models to the use of deep learning-based approaches in the field of video-based human action recognition. The reason has primarily been because of the accessibility, adaptability, accuracy and decrease in the time required to execute the model that deep learning-based approaches offer (Gu et al., 2010; J. Liu et al., 2010; Schüldt et al., 2004; Xu et al., 2017). One of the first significant attempts made was to use two separate stream of CNN's (convolutional neural network) with the aim of training them to extract features from a sampled RGB video frame that was paired with the surrounding stack of optical flow images (Simonyan & Zisserman, 2014). The prediction scores of both streams were fused using averaging and by training a multi-class linear SVM (Crammer & Singer, 2001) on stacked L2-normalized softmax scores as features. There are other works that have followed upon these steps (Donahue et al., 2017; Feichtenhofer et al., 2016; Gkioxari & Malik, 2015; Ng et al., 2015; Wang et al., 2015, 2016). For instance, others built upon the two stream architecture (Simonyan & Zisserman, 2014) in order to create an architecture capable of fusing temporal and spatial cues at several levels of granularity in feature extraction that is also able to integrate spatial and temporal information (Feichtenhofer et al., 2016). More recent work utilized a CNN in order to

learn the optical flow prediction because traditional optical flow methods are computationally expensive and have a burdensome optimization process (Fan et al., 2018; Gao et al., 2018; Hui et al., 2018; Piergiovanni & Ryoo, 2019). Also, another benefit of using CNN's to learn the optical flow prediction is that it reduces the number of trainable parameters that need to be learned as only one network is required. Other methods have explored the benefits of using long short term memory (LSTM)-based structures (Donahue et al., 2017) in order to incorporate motion by updating the pooling of features across time (De Geest & Tuytelaars, 2018; Ge et al., 2019; Perrett & Damen, 2019).

These approaches typically use single video frames as inputs which may be difficult to capture motion actions that are established by longer-horizon temporal correlation. To overcome this, 3D convolutional neural networks were proposed to help learn the spatiotemporal features (Diba et al., 2018; Tran et al., 2015). This is done by using temporally densely sampled sequences of images as inputs. A major downside of the 3D structures is the high number of parameters that are involved in the model and need to be trained for. This leads to an increase in computational cost and necessitates use of large-scale training datasets. To solve this problem an effort has been made to come up with more lightweight and efficient solutions. Some of these works use a (2+1)D decomposition instead of the 3D structure (Diba et al., 2018; He et al., 2019; Li et al., 2020; Qiu et al., 2019; Sun et al., 2015; Tran et al., 2019; Xie et al., 2017) or some works combine 2D CNN and a 3D CNN (J. Lin et al., 2019; Luo & Yuille, 2019; Tran et al., 2018; Xie et al., 2017; Zhou et al., 2018; Zolfaghari et al., 2018).

Even though RGB-video-based algorithms have achieved remarkable results up till now, there still remain several challenging aspects such as background clutter, illumination disparity, viewpoint variation, etc. A possible way that can lead to improvement in recognition performance is to use a skeleton data representation. Earlier models that used the skeleton data representation did not take into account the internal dependencies between body joints and thus dismissed parts of the information related to target action (Du et al., 2015; Fernando et al., 2015; M. Liu et al., 2017; Vemulapalli et al., 2014). Recent works tend to use graph convolutional networks to extract features by creating a skeleton graph which has vertices and edges to represent the joints and bones (Thakkar & Narayanan, 2018; Yan et al., 2018; Yang et al., 2018). These approaches rely on datasets that mainly contain adult motion actions (Carreira & Zisserman, 2017; Damen et al., 2018; Goyal et al., 2017; Gupta et al., 2020; Kuehne et al., 2011; Shahroudy et al., 2016; Soomro et al., 2012).

2.2 Use of Human Action Recognition for Rehabilitation Purposes

Action recognition approaches, such as those described above, can contribute to the field of rehabilitation by being integrated in the automation of assistive devices and assessment of training outcomes. There are several examples of technology applications that offer a wide range of training interaction activities during 'serious games' by giving direct access to an objective performance feedback that utilizes Virtual Reality or camera systems (Alankus et al., 2010; Burke et al., 2009; Collins et al., 2017; Jaume-I-Capó et al., 2014; Jaume-I-Capó & Samčović, 2014). For instance, a user wears a glove or holds onto an object of a single color in order to track the arm

motion (Burke et al., 2009). Another example is the daily activity observation system for stroke patients (Collins et al., 2017). In the latter, depth and skeleton data obtained from a Kinect v2 depth camera are used in order to assess motor actions. Similarly, Leightley et al. analyzed and determined how kinematic data from activities obtained through Kinect sensors can be successfully classified for rehabilitation purposes using Random Forests and Support Vector Machines (Leightley et al., 2013).

Recently, there has been an interesting shift toward pediatric rehabilitation paradigms which involve the use of assistive technology and computer vision approaches. For example the work by Pulido et al. focuses on non-contact upper limb rehabilitation autonomously by utilizing a social robot to perform a set of actions that a child has to imitate (Pulido et al., 2017). The movements performed are captured by a Kinect depth camera and stored as 3D skeletons which are then compared with entries from an existing knowledge base. There are also other applications which make use of multiple camera systems in order to resolve issues caused by occlusions which is very likely to happen in infant rehabilitation sessions (Efthymiou et al., 2018; Kokkoni et al., 2020; Pacheco et al., 2021; Tsiami et al., 2018). An example is the learning environment developed for infants that utilizes socially assistive robots and a body weight support system in which a set of Kinect cameras are used to capture movements and action recognition algorithms are developed to close the loop between the infants and the robots (Kokkoni et al., 2020; Pacheco et al., 2021). Nonetheless, the amount and type of motor actions that have been used in the action recognition studies or research still remains limited.

Chapter 3

Methodology

3.1 Dataset

In this work we use a new annotated dataset on infant reaching recently developed by our team (Dechemi et al., 2021). The dataset provides various descriptors of infant reaching for 17 videos. These descriptors were obtained from the videos through annotation and digitization analyses. In order to collect the videos, YouTube was used. The videos were found by using search terms such as ‘infant’, ‘reaching’, ‘grabbing’, and ‘sitting’ etc. Furthermore, both typically developing infants and atypically developing infants (infants with developmental delays) were considered. The video inclusion criteria were:

1. The video should display an awake and active infant who is between the age ranges of 0-12 months.
2. Subjects should perform at least one reaching action during which the camera remained stationary.
3. Subjects should reach for and complete the reaching movement regardless of the shape and size of the objects.
4. Subjects should perform at least one reaching action during which both hand and object were fully visible throughout the reach.
5. The subject should be in a sitting position during the reaching action.

3.2 Video Annotations

Since the videos were collected from YouTube, each of the videos had a different image resolution and camera placement such as angle, distance from the infant, etc. Thus, it was important to select a known measure as a reference point so that each video could be calibrated uniformly in order to account for these differences. The infant's cornea was found to be the most suitable point of reference across all videos. A horizontal line, drawn from one end of the cornea to the other, was calibrated at 1.05 cm which is the average iris diameter reported for this age range (Ronneburger et al., 2006).

The annotations protocol used for the dataset was divided into distinct phases. The first phase focuses on the annotation of the reaching actions performed by the infants in each of the videos. The second phase provides further details about the reaching action such as the hand selection performing the reaching action, position of the hand with respect to the body at RN, position of the presented object with respect to the body and angle of the camera used in the recording. The details of the analysis in the second phase are provided in Table 1.

This thesis considers the annotations conducted in the first phase. In more detail, during the first phase the Reaching Onset (RN) and the Reaching Offset (RF) for every reaching action in each video were annotated. The Reaching Onset was defined as the first frame in which the infant's hand (left, right, or both) began moving toward the presented object and this movement continuously occurred for at least 5 frames. The Reaching Offset was defined as the first frame in which the infant's hand

intentionally touched the object (Figure 1). Segments where an object was transferred from one hand to the other hand, partial reaching actions (that is, reaching actions that were interrupted or unfinished), and also actions in which the hand and/or object were occluded were excluded from further analyses because it would not be possible to make an accurate prediction of the Reaching Offset without severely hampering the annotation reliability. Furthermore, in order to ensure that the network had sufficient data for action detection only reaching actions that were greater than six frames in length were considered. In addition to this, frames within 3 seconds before RN and 3 seconds after RF from video segments that contained the selected actions were cropped and labelled as no reaching. These frames were used for further analyses and for distinguishing between a reaching action and a non-reaching action.

Figure 1. Examples of Reaching Onset (left) and Offset (Right) Frames

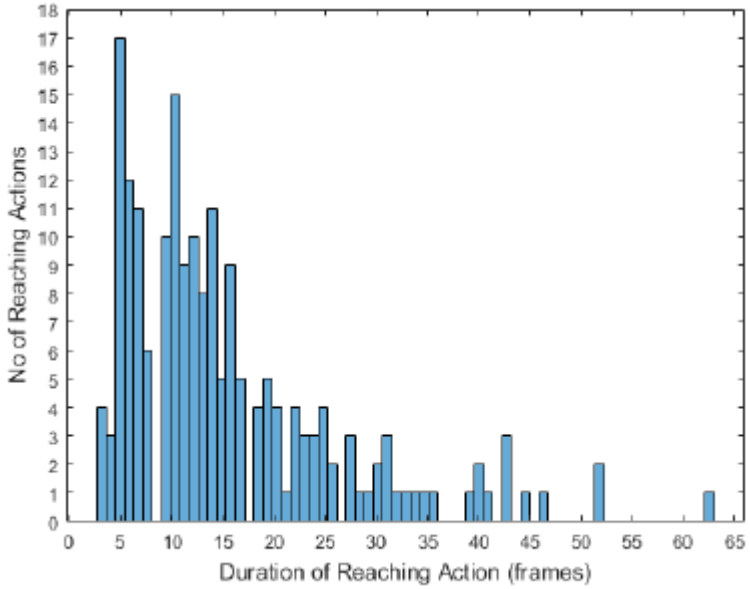


Figure 2**. Distribution of Reaching Actions**

To ensure the annotations were reliable, a reliability protocol was established. Two sample videos were initially used to determine the reliability of the four annotators. Reaching actions from these two sample videos were annotated by all the annotators and only when they all achieved a 100% agreement on the frequency of reaching actions and a +/-3 frame selection difference between RN and RF, they proceeded with analysis of the remaining videos (two annotators assigned per remaining video). After the annotation process the dataset included a total of 193 reaches with 86 reaches being performed by the left hand and 107 being performed by the right one. A total of 57 different objects were involved in the reaching actions (the details are provided in table 2). Also, Figure 2 shows the distribution of the duration of the reaches in the dataset and provides a glimpse into the large variability of reaching action that is encountered.

Table 1: Categories and Subcategories Annotated to Describe a Reaching Action

Stage of Reaching Action	Hand used to Reach	Hand's Initial Position in Relation to Body	Object's Position in Relation to Body (Vertical)	Object's Position in Relation to Body (Horizontal)	Camera View
Onset (RN)	Left	Above Head	Above Head	To the Left	Top
Offset (RF)	Right	Head Level	Head Level	In Front	Front
		Chest Level	Chest Level	To the Right	Side
		Hip Level	Hip Level		
		Floor	Floor		

Table 2: Subject Information and Corresponding No. of Reaches and Objects Annotated

Subject ID	Age [months]	Gender [M / F]	# Reaches LH	RH	Total	Total Obj.
T01	6-8	F	5	3	8	4
T02	8-10	M	8	4	12	4
T03	11-12*	M	2	8	10	2
T04	6-12	F	3	5	8	5
T05	10-12	M*	3	1	4	2
T06	10-12	M*	0	2	2	2
T07	6-7*	M	2	3	5	3
T08	<12*	M	4	5	9	1
T09	6	F	3	6	9	2
T10	6	F	3	3	6	3
T11	6-8	M	2	1	3	2
T12	8	F	16	33	49	8
T13	10	F	5	9	14	5
T14	6	F	1	1	2	2
T15	9	M	20	20	40	10
T16	7	F	2	1	3	2
D01	6-9	F	1	-	1	1
D02	10	M	-	1	1	1
D03	<12*	M	4	1	5	1
D04	9-12	M	1	-	1	1
D05	<12*	M	1	-	1	1
Total	–	–	**86**	**107**	**193**	**57**

3.3 Bounding Box Annotations

The bounding boxes were used to detect each of the hands (LH, RH) and any object in each of the frames. The primary purpose of doing this was to assess the spatial and temporal connection that is present during reaching. A total of 607 frames were sampled randomly from the total 2,984 frames in the dataset by two researchers in the team in order to detect the infant, the right and left hand and the objects involved in the reaching action.

The sampled images and the corresponding bounding boxes were used to train an object detector that helped in automating the process for obtaining the bounding boxes for the remaining frames/images. In order to train the object detector a Yolo-v3 detector (Redmon & Farhadi, 2018) pre-trained on the COCO dataset (T. Y. Lin et al., 2014) was fine-tuned using the obtained annotations. During fine-tuning the detector was trained to recognize four classes: 1) Infant; 2) Left Hand; 3) Right hand; and 4) Object.

The object detector was trained on 75% of the data and tested on 25% of the data. Figure 3 provides some sample illustrative examples of infants in different environments.

Figure 3: Sample Frames from the Dataset

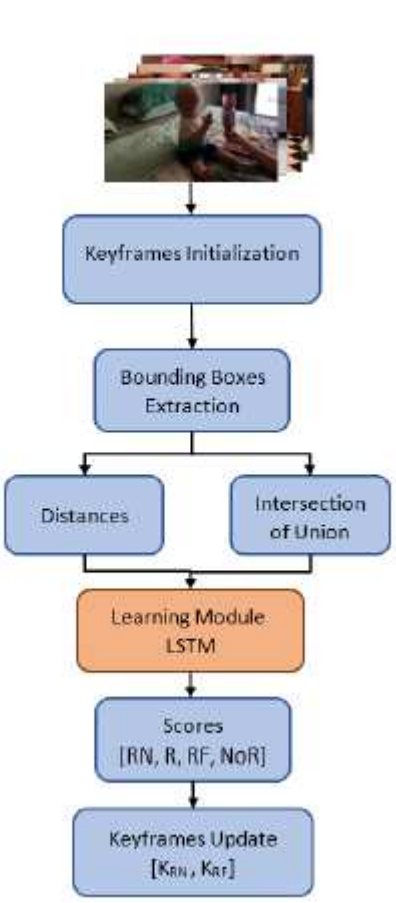
3.4 Infant Reaching Action Algorithm

The proposed approach for the infant reaching action recognition is depicted in Figure 4. In order to validate the proposed algorithm, the RN and RF annotations obtained through the analysis in section 3.2 were utilized at various stages.

The objects, the left hand and the right hand in a video containing a sequence of T video frames is denoted using $X = \{x_i^T\}$ and $H = \{h_L^T, h_R^T\}$ respectively in the frame sequence. It is to be noted that in order to find the precise location of the objects and the hands in the images, no tracking information is incorporated and the algorithm only relies on the coordinates of the centers of the bounding boxes as the main source of information. The coordinates of the bounding boxes are acquired through object detection along with the labels for each frame T_i. The two keyframes, T_{RN} for the onset and T_{RF} for the offset, are initialized at the first frame. The reaching action is divided into two distinct phases' onset and offset in order to enable the

spatiotemporal reasoning between X and H, where each phase is detected through a separate process.

Figure 4: Approach Used in the Infant Reaching Action Algorithm



3.4.1 The Reaching Onset Phase

First, the distance d^i_j between the hands $\{h_L^T, h_R^T\}$ and the object x_i in X for the current frame was computed by using the bounding box co-ordinates. Next the distance in the current frame d^i_j was compared with the distance in the previous frame d^i_{j-1}. If $d^i_j - d^i_{j-1} < 0$, then the onset detection is still valid and T_{RN} is kept as

a key frame else if it is noticed that $d^i_j - d^i_{j-1} > 0$ and this increase is seen in consecutive frames, then the onset is invalidated and T_i is set as the new key frame T_{RN}. The reason for employing this strategy is that it helps avoid false negatives during the search process.

3.4.2 Reaching Offset Phase

The first step in determining the Reaching Offset key frame is to estimate the intersection over union (IOU) between the hands $\{h_L^T, h_R^T\}$ and the object x_i. The IOU is estimated and compared to a threshold value that is determined through a learning process. If the IOU value is less than the determined threshold, then no offset is detected and the key frame T_{RF} is definitively set as T_i. If the IOU value is greater than or equal to the determined threshold then an offset is detected and T_{RF} is kept as the key frame.

3.4.3 BabyNet Network Structure

The backbone of the network is a LSTM structure that learns the correlation of reaching between the two bounding boxes for the object and hand through the input which consists of the distances and the IOU's. The final output gives the scores for the onset (RN), the offset (RF), reach (R) and no reach (NoR) which are used to update the identified key frames. The label reach (R) is used for all the frames that fall between the onset RN and offset RF. Similarly, the label no reach or NoR is used for all the frames that come before the onset RN and after the offset RF.

Chapter 4

Experiments

In this section, several ablation experiments were conducted to test the effectiveness of different models on the dataset. For this study, five baseline networks were considered for comparison. The five networks trained are outlined below and the comparison results, including the classification accuracies and inference protocols, are shown in Table 3.

4.1 ResNet

This network was developed by the Facebook AI research team and is a 50 layer deep neural network. Traditionally deep neural networks have been difficult to train. However, this network introduced a residual learning framework that eased the training of substantially deep neural networks. It overcame the problem of Vanishing Gradient which arises as the networks are made deep. ResNet allows successful training of such deep networks by constructing the network through modules called residual modules. The residual module creates a direct path between the input and output to the module by implying an identity mapping. So, the added layer just needs to learn the features on top of the already available input.

Since, the size of the dataset was relatively small, training the dataset on the entire network would overfit the data and lead to poor results. Therefore, in order to avoid this problem a pre-trained ResNet-50 model that has already been trained on the ImageNet dataset was used. Only the last Bottleneck block in the fourth layer was retrained along with the fully connected layer. This method of training achieved an

average training accuracy of 94.59% and an average validation accuracy of 53.65%. However, for further verification the trained model was used to make predictions on the test video and the results indicated that even the slightest movement by the infant or the movement of the adult's hand was classified as a reach.

4.2 ResNet + LSTM

In order to alleviate the problem of the ResNet model misclassifying the reach based on the slightest movement by the infant or the movement of the therapists' hand, we decided to integrate an LSTM block after the average pooling layer of the final residual block of a pre-trained ResNet-50 model. This was done with the aim of learning both the spatio and temporal features of both the reach and no reach movements. After retraining the model, an average training accuracy 0f 94.31% and an average validation accuracy of 54.53% is achieved. Even though the test accuracy was lower for the ResNet+LSTM model as compared to the ResNet model. The one advantage that the model with the LSTM had was that it was successfully able to identify the onset of the reach movement; however, it wasn't able to identify the offset thus leading to lower test accuracy. In addition to this, there was less fluctuation between the reach and no reach predictions compared to the prior network. This indicates that learning the temporal correlation between the frames helped in reducing the number of false negatives for the reaches.

4.3 3D Networks

Since, the ResNet+LSTM model did not achieve the desired result but it showed an improvement over the base ResNet model by learning the temporal features. We

decided to train certain 3D-CNN networks as they incorporate the frames as an extra dimension when performing the convolution operation, thereby learning the temporal features of the entire action as compared to only a few frames with the LSTM.

In order to utilize our dataset for training a 3D-CNN network, the first step was to ensure that the two action classes have the same number of frames for the duration of the action. This was done by manipulating the individual frame rate or FPS (frames per second) for each of the videos. The next challenge was fixing the number of frames for each of the videos. To overcome this challenge, we plotted a histogram showing the frequency and the duration of the reaches shown in figure 2. The histogram tells us that the largest number of frames in a video is 63. Therefore, we decided to upsample the number of frames for all the videos to this number in order to avoid loss of frames for the reach videos. Given the small size of the dataset we decided to create our own custom 3D-CNN model which contained two 3D convolution blocks followed by three fully connected layers. The reason for using three fully connected was to gradually reduce the output size from [1,512] to [1,2] in order to prevent excessive loss of useful information. Furthermore, in order to gauge the performance of our custom 3D-CNN network, we compared its performance with the C3D network (Tran et al., 2015) which is a benchmark 3D-CNN network.

4.4 LSTM with Optical Flow (O-LSTM)

As we detail in Chapter 5 below, the large networks considered up to this point (Sections 4.1-4.3) cannot predict the actions to the desired level of accuracy because they solely rely on visual content or cues. Therefore, it is very important to include

information about the motion in the video sequence. To do this we extracted the optical flow images for the two sets of actions from their respective videos. The optical flow images were obtained using the Farneback method (Farnebäck, 2003). These extracted images were used to train a LSTM model. Before the model can be trained, the dimension of the image inputs needs to be flattened to a size of 1x12,288 which is obtained from a reduced size of 64x64 source image. This was primarily done to reduce the training time and also reduce the number of trainable parameters. The LSTM model trained contained a single layer with 50% dropout.

4.5 Models Trained Using Bounding Box Coordinates

A Multi-Layer Perceptron (MLP) network is trained along with the BabyNet proposed in this work in order to compare their performance. Both models utilize the center coordinates of the bounding boxes in order to gauge the position of the hands and object in an image. The MLP network is trained with a single hidden layer and uses two inputs to generate four outputs.

Chapter 5

Results

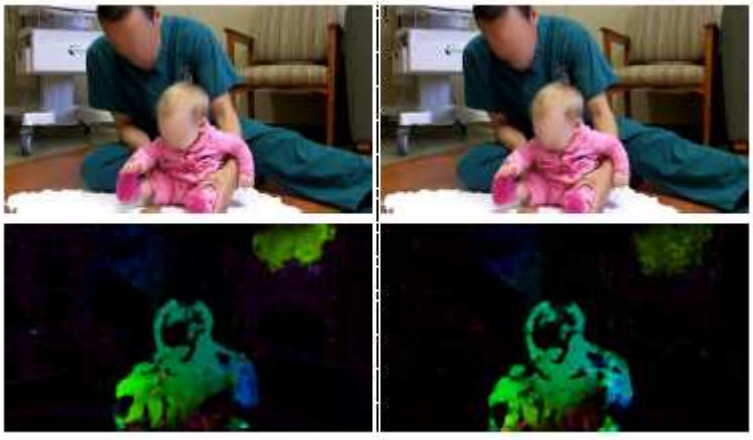
5.1 Effect of Data Augmentation

Firstly, it can be seen from the first component of Table 3 that the ResNet model with data augmentation has a better average test accuracy than the ResNet model without data augmentation (58.16% vs 53.43%). Next, to properly gauge the effect of data augmentation on the accuracy, we need to compare the tradeoff between precision and recall of the two networks. It can be seen from Table 3 that the ResNet model with data augmentation has a higher Precision and recall score. This indicates that the False Positive rate for the network with data augmentation is lower which is better as the network is triggered fewer times.

Table 3: Evaluation of the Models Trained on the Infant Reaching Dataset

Model	Parameters (Trainable)/(Total)	Avg. Training Accuracy [%]	Avg. Validation Accuracy [%]	Avg. Testing Accuracy [%]	Precision NoR / R	Recall NoR / R
ResNet No Data Augm.	4,468,739 / 23,514,179	98.30	50.61	53.43	0.65 / 0.40	0.55 / 0.51
ResNet With Data Augm.	4,468,739 / 23,514,179	94.59	53.65	58.16	0.68 / 0.45	0.62 / 0.52
ResNet+LSTM No Data Augm.	9,186,819 / 28,232,259	98.61	50.59	54.21	0.65 / 0.41	0.57 / 0.49
ResNet+LSTM With Data Augm.	9,186,819 / 28,232,259	94.31	54.53	54.42	0.68 / 0.42	0.51 / 0.60
O-LSTM	117,460,994 / 117,460,994	75.44	81.16	63.71	0.59 / 0.82	0.92 / 0.35
MLP	144 / 144	47.66	46.13	51.8	0.55 / 0.67	0.78 / 0.42
BabyNet (Ours)	1,204 / 1,204	44.45	38.93	66.27	0.57 / 0.66	0.72 / 0.49

Figure 5: The Images on the Top Depict the RGB Images and the Images on the Bottom Show the Equivalent Optical Flow Images



5.2 Effect of Integrating the LSTM Block

Integrating a LSTM block to the baseline ResNet model did not have any significant improvement in the performance and actually produced similar results as the standalone ResNet models. The average testing accuracy was similar for both the architectures and in fact decreased for the ResNet+LSTM model with data augmentation compared to the ResNet model with data augmentation from 58.16% to 54.42%. The similarity in performance is further highlighted by the precision scores as both the networks have the same precision for the no reaches and a minor difference for the reaches (0.45 for ResNet to 0.42 for ResNet+LSTM). However, a significant difference is noticed in the recall score as it increased for the reaches from 0.51 to 0.60 thus showing that the ResNet+LSTM model produces a lower number of false negatives.

5.3 Performance of the 3D Networks

Upon training and evaluating the two 3D baseline networks, it is evident that the custom 3D network designed by us outperforms the C3D network for our dataset as shown in Table 4 It is to be noted that the 3D networks performed well in classifying the videos of the actions individually. However, when tested on the test video which contained both actions, the network did not predict any of the reaches successfully. The 2D networks outperformed the 3D networks in this aspect as they were able to predict part of the reach actions correctly on the test video.

Table 4: Evaluation of the 3D Networks Trained on the Infant Reaching Dataset

Model	Parameters (Trainable) / (Total)	Avg. Training Accuracy [%]	Avg. Validation Accuracy [%]	Avg. Testing Accuracy [%]
3D-CNN	202,394,946 / 202,394,946	89.64	58.33	80
C3D	222,597,351 / 222,597,351	55.85	70	70

5.4 Comparison of BabyNet and the MLP Network

It can be seen from Table 4 that the BabyNet outperforms the MLP based on the test accuracy for the same split (66.27 vs 51.8). Furthermore, both networks predicted the same number of frames incorrectly during the reach but the MLP predicted 20 frames incorrectly during the no reach action whereas the BabyNet only predicted 6 frames incorrectly. However, the BabyNet only had a delay of 1 frame while predicting the reach whilst the MLP had a delay of 4 frames. Overall, the performance of the

BabyNet was better than the MLP due to a smaller delay while predicting a reach and making fewer incorrect predictions.

5.5 Performance of the O-LSTM Network

Results show that the O-LSTM trained on the optical flow images achieves an average testing accuracy of 63.71% which is higher than all the ResNet-based architectures. Upon further analysis of the precision and recall scores for both the reaches and no reaches, it is seen that the model has lower false positive detection for the reaches compared to that of no reaches (0.59 to 0.82). However, the detections are more likely to be false negatives for the reaches given that the recall score is only 0.35. Another observation that is made is that it is a challenge for the network to recognize short reaches. In addition to this, the O-LSTM requires a significantly larger number of parameters to be trained as the input images need to be flattened before training thus causing the size to be transformed from 64x64 to 1x12288. The large number of trainable parameters in addition to the high computational effort required during transforming the original images to the optical flow images suggest that the O-LSTM network is not the most efficient as compared to the BabyNet. Figure 5 provides a few sample optical flow images.

Conclusion

This work proposes a new lightweight network, BabyNet, that is able to model both short- and long-range motion correlation related to different key phases of a reaching action. A new dataset suitable for infant action recognition is used and the performance of the BabyNet and several state-of-the-art structures is evaluated on this dataset.

Results indicate that BabyNet is small in structure but powerful and can challenge significantly larger structures by achieving an average testing accuracy of 66.27% on the proposed dataset. Upon evaluation of the ResNet based structures, an increase in the rate of false positives is seen despite solid performance in terms of the training and validation accuracy. The O-LSTM structure had the second best score in terms of the testing accuracy but could not correctly balance between the scores of recall (0.92) and precision (0.35). However, it still remains a worthy approach for further consideration as the optical flow images can better distinguish subtle motion patterns compared to RGB images.

Another network tested, MLP, is of comparable size to the BabyNet but did not perform as well as BabyNet. BabyNet had an approximate improvement of 27% in performance compared to the MLP with almost the same precision and recall scores. The key differentiator between the two networks was that BabyNet provided the onset and offset frames at a delay of one frame whereas the MLP had a delay of four frames. These finding indicate that BabyNet is capable of serving as a lightweight network for the task of video-based infant action recognition.

References:

Alankus, G., Lazar, A., May, M., & Kelleher, C. (2010). Towards customizable games for stroke rehabilitation. *Conference on Human Factors in Computing Systems - Proceedings*, *3*, 2113–2122. https://doi.org/10.1145/1753326.1753649

Arnold, A. J., Haworth, J. L., Moran, V. O., Abulhasan, A., Steinbuch, N., & Kokkoni, E. (2020). Exploring the Unmet Need for Technology to Promote Motor Ability in Children Younger Than 5 Years of Age: A Systematic Review. *Archives of Rehabilitation Research and Clinical Translation*, *2*(2), 100051. https://doi.org/10.1016/j.arrct.2020.100051

Berthier, N. E., & Keen, R. (2006). Development of reaching in infancy. *Experimental Brain Research*, *169*(4), 507–518. https://doi.org/10.1007/s00221-005-0169-9

Burke, J. W., McNeill, M. D. J., Charles, D. K., Morrow, P. J., Crosbie, J. H., & McDonough, S. M. (2009). Serious games for upper limb rehabilitation following stroke. *Proceedings of the 2009 Conference in Games and Virtual Worlds for Serious Applications, VS-GAMES 2009*, 103–110. https://doi.org/10.1109/VS-GAMES.2009.17

Campos, J. J., Anderson, D. I., Barbu-Roth, M. A., Hubbard, E. M., Hertenstein, M. J., & Witherington, D. (2000). Travel Broadens the Mind. *Infancy*, *1*(2), 149–219. https://doi.org/10.1207/S15327078IN0102_1

Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*, 4724–4733. https://doi.org/10.1109/CVPR.2017.502

Centers for Disease Control and Prevention (2019). (2014). Surveillance Report of Traumatic Brain Injury-related Emergency Department Visits, Hospitalizations, and Deaths. *Centers for Disease Control and Prevention, U.S. Department of Health and Human Services*.

Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S. R., Bogen, D. K., Prosser, L., Johnson, M. J., & Kording, K. P. (2020). Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(11), 2431–2442. https://doi.org/10.1109/TNSRE.2020.3029121

Collins, J., Warren, J., Ma, M., Proffitt, R., & Skubic, M. (2017). Stroke patient daily activity observation system. *Proceedings - 2017 IEEE International Conference on*

Bioinformatics and Biomedicine, BIBM 2017, *2017-January*, 844–848. https://doi.org/10.1109/BIBM.2017.8217765

Crammer, K., & Singer, Y. (2001). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. In *Journal of Machine Learning Research* (Vol. 2). https://doi.org/10.5555/944790.944813

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2018). *Scaling Egocentric Vision: The EPIC-KITCHENS Dataset*. http://youtu.be/Dj6Y3H0ubDw.

De Geest, R., & Tuytelaars, T. (2018). Modeling temporal structure with LSTM for online action detection. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-January*, 1549–1557. https://doi.org/10.1109/WACV.2018.00173

Dechemi, A., Bhakri, V., Sahin, I., Modi, A., Mestas, J., Peiris, P., Barrundia, D., Kokkoni, E., & Karydis, K. (2021). BabyNet: A Lightweight Network for Infant Reaching Action Recognition in Unconstrained Environments to Support Pediatric Rehabilitation Applications(Under Review). *2021 IEEE RO-MAN*.

Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2018). Temporal 3D ConvNets using Temporal Transition Layer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1117–1121.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 677–691. https://doi.org/10.1109/TPAMI.2016.2599174

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714

Efthymiou, N., Koutras, P., Filntisis, P. P., Potamianos, G., & Maragos, P. (2018). Multi-View Fusion for Action Recognition in Child-Robot Interaction. *Proceedings - International Conference on Image Processing, ICIP*, 455–459. https://doi.org/10.1109/ICIP.2018.8451146

Emeli, V., Fry, K. E., & Howard, A. (2020). Towards Infant Kick Quality Detection to Support Physical Therapy and Early Detection of Cerebral Palsy: A Pilot Study. *29th IEEE International Conference on Robot and Human Interactive*

Communication, RO-MAN 2020, 1069–1074. https://doi.org/10.1109/RO-MAN47096.2020.9223571

Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., & Huang, J. (2018). End-to-End Learning of Motion Representation for Video Understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6016–6025. http://arxiv.org/abs/1804.00413

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2749*, 363–370. https://doi.org/10.1007/3-540-45103-x_50

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 1933–1941. https://doi.org/10.1109/CVPR.2016.213

Fernando, B., Gavves, E., José Oramas, M., Ghodrati, A., & Tuytelaars, T. (2015). Modeling video evolution for action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 5378–5387. https://doi.org/10.1109/CVPR.2015.7299176

Fetters, L. (2010). Perspective on variability in the development of human action. *Physical Therapy*, *90*(12), 1860–1867. https://doi.org/10.2522/ptj.2010090

Gao, R., Xiong, B., & Grauman, K. (2018). Im2Flow: Motion Hallucination from Static Images for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 5937–5947. https://doi.org/10.1109/CVPR.2018.00622

Ge, H., Yan, Z., Yu, W., & Sun, L. (2019). An attention mechanism based convolutional LSTM network for video action recognition. *Multimedia Tools and Applications*, *78*(14), 20533–20556. https://doi.org/10.1007/s11042-019-7404-z

Gkioxari, G., & Malik, J. (2015). Finding action tubes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 759–768. https://doi.org/10.1109/CVPR.2015.7298676

Goto, J., Kidokoro, T., Ogura, T., & Suzuki, S. (2013). Activity recognition system for watching over infant children. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 473–477. https://doi.org/10.1109/ROMAN.2013.6628549

Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., & Memisevic, R. (2017). The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-October*, 5843–5851. https://doi.org/10.1109/ICCV.2017.622

Gu, J., Ding, X., Wang, S., & Wu, Y. (2010). Action and gait recognition from recovered 3-D human joints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *40*(4), 1021–1033. https://doi.org/10.1109/TSMCB.2010.2043526

Gupta, P., Thatipelli, A., Aggarwal, A., Maheshwari, S., Trivedi, N., Das, S., & Sarvadevabhatla, R. K. (2020). Quo Vadis, Skeleton Action Recognition ? *ArXiv*. http://arxiv.org/abs/2007.02072

Hadfield, J., Chalvatzaki, G., Koutras, P., Khamassi, M., Tzafestas, C. S., & Maragos, P. (2019). A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task. *IEEE International Conference on Intelligent Robots and Systems*, 1251–1256. https://doi.org/10.1109/IROS40897.2019.8968443

Hashemi, J., Tepper, M., Vallin Spina, T., Esler, A., Morellas, V., Papanikolopoulos, N., Egger, H., Dawson, G., & Sapiro, G. (2014). Computer Vision Tools for Low-Cost and Noninvasive Measurement of Autism-Related Behaviors in Infants. *Autism Research and Treatment*, *2014*, 1–12. https://doi.org/10.1155/2014/935686

He, D., Zhou, Z., Gan, C., Li, F., Liu, X., Li, Y., Wang, L., & Wen, S. (2019). StNet: Local and global spatial-temporal modeling for action recognition. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, *33*(01), 8401–8408. https://doi.org/10.1609/aaai.v33i01.33018401

Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U. G., Weinberger, R., & Sebastian Schroeder, A. (2019). Computer vision for medical infant motion analysis: State of the art and RGB-D data set. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11134 LNCS*, 32–49. https://doi.org/10.1007/978-3-030-11024-6_3

Hui, T. W., Tang, X., & Loy, C. C. (2018). LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8981–8989. https://doi.org/10.1109/CVPR.2018.00936

Jaume-I-Capó, A., Martinez-Bueso, P., Moya-Alcover, B., & Varona, J. (2014). Interactive rehabilitation system for improvement of balance therapies in people with cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(2), 419–427. https://doi.org/10.1109/TNSRE.2013.2279155

Jaume-I-Capó, A., & Samčović, A. (2014). Vision-based interaction as an input of serious game for motor rehabilitation. *2014 22nd Telecommunications Forum, TELFOR 2014 - Proceedings of Papers*, 854–857. https://doi.org/10.1109/TELFOR.2014.7034540

Kokkoni, E., Mavroudi, E., Zehfroosh, A., Galloway, J. C., Vidal, R., Heinz, J., & Tanner, H. G. (2020). GEARing smart environments for pediatric motor rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *17*(1), 1–15. https://doi.org/10.1186/s12984-020-0647-0

Konczak, J., & Dichgans, J. (1997). The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental Brain Research*, *117*(2), 346–354. https://doi.org/10.1007/s002210050228

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543

Leightley, D., Darby, J., Li, B., Mcphee, J. S., & Yap, M. H. (2013). Human activity recognition for physical rehabilitation. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 261–266. https://doi.org/10.1109/SMC.2013.51

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., & Wang, L. (2020). TEA: Temporal Excitation and Aggregation for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 906–915. https://doi.org/10.1109/CVPR42600.2020.00099

Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-October*, 7082–7092. https://doi.org/10.1109/ICCV.2019.00718

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, J., Yang, J., Zhang, Y., & He, X. (2010). Action recognition by multiple features and hyper-sphere multi-class SVM. *Proceedings - International Conference on Pattern Recognition*, 3744–3747. https://doi.org/10.1109/ICPR.2010.912

Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, *68*, 346–362. https://doi.org/10.1016/j.patcog.2017.02.030

Luo, C., & Yuille, A. (2019). Grouped spatial-temporal aggregation for efficient action recognition. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-October*, 5511–5520. https://doi.org/10.1109/ICCV.2019.00561

Marinoiu, E., Zanfir, M., Olaru, V., & Sminchisescu, C. (2018). 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2158–2167. https://doi.org/10.1109/CVPR.2018.00230

McCay, K. D., Ho, E. S. L., Shum, H. P. H., Fehringer, G., Marcroft, C., & Embleton, N. D. (2020). Abnormal Infant Movements Classification with Deep Learning on Pose-Based Features. *IEEE Access*, *8*, 51582–51592. https://doi.org/10.1109/ACCESS.2020.2980269

Nectac. (2011). *The Importance of Early Intervention for Infants andToddlers with Disabilities and their Families*. https://doi.org/10.1111/j.1469

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101

Nie, Q., Wang, X., Wang, J., Wang, M., & Liu, Y. (2018). A child caring robot for the dangerous behavior detection based on the object recognition and human action recognition. *2018 IEEE International Conference on Robotics and Biomimetics, ROBIO 2018*, 1921–1926. https://doi.org/10.1109/ROBIO.2018.8665218

Pacheco, C., Mavroudi, E., Kokkoni, E., Tanner, H. G., & Vidal, R. (2021). A Detection-based Approach to Multiview Action Classification in Infants. *2020 25th International Conference on Pattern Recognition (ICPR)*, 6112–6119. https://doi.org/10.1109/ICPR48806.2021.9412822

Perrett, T., & Damen, Di. (2019). DDLSTM: Dual-domain LSTM for cross-dataset action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 7844–7853.

https://doi.org/10.1109/CVPR.2019.00804

Piergiovanni, A. J., & Ryoo, M. S. (2019). Representation flow for action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 9937–9945. https://doi.org/10.1109/CVPR.2019.01018

Pulido, J. C., González, J. C., Suárez-Mejías, C., Bandera, A., Bustos, P., & Fernández, F. (2017). Evaluating the Child–Robot Interaction of the NAOTherapist Platform in Pediatric Rehabilitation. *International Journal of Social Robotics*, *9*(3), 343–358. https://doi.org/10.1007/s12369-017-0402-2

Qiu, Z., Yao, T., Ngo, C. W., Tian, X., & Mei, T. (2019). Learning spatio-temporal representation with local and global diffusion. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 12048–12057. https://doi.org/10.1109/CVPR.2019.01233

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. http://arxiv.org/abs/1804.02767

Ronneburger, A., Basarab, J., & Howland, H. C. (2006). Growth of the cornea from infancy to adolescence. *Ophthalmic and Physiological Optics*, *26*(1), 80–87. https://doi.org/10.1111/j.1475-1313.2005.00362.x

Schneiberg, S., Sveistrup, H., McFadyen, B., McKinley, P., & Levin, M. F. (2002). The development of coordination for reach-to-grasp movements in children. *Experimental Brain Research*, *146*(2), 142–154. https://doi.org/10.1007/s00221-002-1156-z

Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. *Proceedings - International Conference on Pattern Recognition*, *3*, 32–36. https://doi.org/10.1109/ICPR.2004.1334462

Sciortino, G., Farinella, G. M., Battiato, S., Leo, M., & Distante, C. (2017). On the Estimation of Children's Poses. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10485 LNCS*, 410–421. https://doi.org/10.1007/978-3-319-68548-9_38

Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 1010–1019. https://doi.org/10.1109/CVPR.2016.115

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, *1*(January), 568–576. http://arxiv.org/abs/1406.2199

Soomro, K., Zamir, A. R., & Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. http://arxiv.org/abs/1212.0402

Stahl, A., Schellewald, C., Stavdahl, Ø., Aamo, O. M., Adde, L., & Kirkerod, H. (2012). An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *20*(4), 605–614. https://doi.org/10.1109/TNSRE.2012.2195030

Sun, L., Jia, K., Yeung, D.-Y., & Shi, B. E. (2015). Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4597–4605.

Suzuki, S., Mitsukura, Y., Igarashi, H., Kobayashi, H., & Harashima, F. (2012). Activity recognition for children using self-organizing map. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 653–658. https://doi.org/10.1109/ROMAN.2012.6343825

Taylor, C. A., Bell, J. M., Breiding, M. J., & Xu, L. (2017). Traumatic brain injury-related emergency department visits, hospitalizations, and deaths - United States, 2007 and 2013. *MMWR Surveillance Summaries*, *66*(9), 1–16. https://doi.org/10.15585/mmwr.ss6609a1

Thakkar, K., & Narayanan, P. J. (2018). Part-based Graph Convolutional Network for Action Recognition. *ArXiv*. http://arxiv.org/abs/1809.04983

Tran, D., Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *IN ICCV*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.714.5752

Tran, D., Wang, H., Feiszli, M., & Torresani, L. (2019). Video classification with channel-separated convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-October*, 5551–5560. https://doi.org/10.1109/ICCV.2019.00565

Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., & Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

Tsiami, A., Koutras, P., Efthymiou, N., Filntisis, P. P., Potamianos, G., & Maragos, P.

(2018). Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. *Proceedings - IEEE International Conference on Robotics and Automation*, 4585–4592. https://doi.org/10.1109/ICRA.2018.8461210

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3D skeletons as points in a lie group. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 588–595. https://doi.org/10.1109/CVPR.2014.82

Wang, L., Xiong, Y., Wang, Z., & Qiao, Y. (2015). *Towards Good Practices for Very Deep Two-Stream ConvNets*. http://arxiv.org/abs/1507.02159

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9912 LNCS*, 20–36. http://arxiv.org/abs/1608.00859

Westeyn, T. L., Abowd, G. D., Starner, T. E., Johnson, J. M., Presti, P. W., & Weaver, K. A. (2012). Monitoring children's developmental progress using augmented toys and activity recognition. *Personal and Ubiquitous Computing*, *16*(2), 169–191. https://doi.org/10.1007/s00779-011-0386-0

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2017). Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11219 LNCS*, 318–335. http://arxiv.org/abs/1712.04851

Xu, D., Xiao, X., Wang, X., & Wang, J. (2017). Human action recognition based on Kinect and PSO-SVM by representing 3D skeletons as points in lie group. *ICALIP 2016 - 2016 International Conference on Audio, Language and Image Processing - Proceedings*, 568–573. https://doi.org/10.1109/ICALIP.2016.7846646

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 7444–7452. http://arxiv.org/abs/1801.07455

Yang, Z., Li, Y., Yang, J., & Luo, J. (2018). Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2018.2864148

Zhou, Y., Sun, X., Zha, Z. J., & Zeng, W. (2018). MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition. *Proceedings of the IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, 449–458. https://doi.org/10.1109/CVPR.2018.00054

Zolfaghari, M., Singh, K., & Brox, T. (2018). ECO: Efficient Convolutional Network for Online Video Understanding. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11206 LNCS*, 713–730. http://arxiv.org/abs/1804.09066