Video-Based Sports Activity Recognition for Children

Feyisayo Olalere*, Vincent Brouwers*, Metehan Doyran*, Ronald Poppe* and Albert Ali Salah* †
* Utrecht University, Utrecht, the Netherlands
E-mail: m.doyran@uu.nl
† Boğaziçi University, Istanbul, Turkey

***Abstract*—Large-scale action recognition datasets contain more instances of adults than children, and models trained with these datasets may not perform well for children. In this study, we test if current state-of-the-art deep learning models have some systemic bias in decoding the activity being performed by an adult or a child. We collected a sports activity recognition dataset with child and adult labels. We fine-tuned a state-of-the-art action recognition classifier on two different segments of our dataset, containing only children or only adults. Our results show that cross-condition generalization performance of the resulting networks is not similar. Our results indicate that the child-specific segment is more complex to generalize than the adult-specific segment. The dataset and the code are made publicly available[1].**

I. Introduction

Activity recognition for human behavior analysis is one of the major problems being researched within the computer vision community [1]. The increase of interest in this research area is fueled by the availability of more datasets, increased hardware complexity, advanced computer vision techniques, and the need for various applications in the real world [2]. These applications include video surveillance systems, robotics for human behavior characterization, medical diagnosis, and many more [3], [4]. Within this area, child activity analysis is an important problem with several high-impact application scenarios.

Cognitive and neuromotor development of children can be affected from disorders such as cerebral palsy, muscular dystrophy and other neuromotor disorders. Early diagnosis of these disorders can be done with the aid of computational approaches [5], [6]. Visual analysis of children can provide important cues about development [7], [8]. There are several efforts to collect video recordings of infants during interactions, such as the PLAY project[2], which aims to "collect, code, and share 900 hours of video collected in the homes of children at 12, 18, and 24 months of age drawn from 30 sites across North America." Computer vision based automatic analysis of child behavior, especially the analysis of affective cues, are considered for psychotherapy and children's play therapy [9], [10], and for analysis of gaming interactions [11].

Publicly available tools to analyse children's behavior and actions can aid medical experts in tracking and quantifying such behavior objectively [12]. However the current activity recognition field (both datasets and methods) is heavily skewed towards adults. We hypothesize that current activity recognition datasets, as they include small numbers of children compared to adults, are not suitable to train deep neural networks with enough generalizing power to create powerful analysis tools targeting child behavior. In this paper, we test this hypothesis by building a special database, and by investigating the performance of state of the art activity recognition approaches under different training conditions.

For the activity analysis application, we choose sports activities as our focus domain because there are many publicly accessible multimedia resources for this problem and the sports domain provides rich body configurations for analysis.

Many of the current state of the art (SOTA) deep learning models are trained with more adult data than child data. When we look at one of the largest activity recognition datasets, i.e. Kinetics-400 [13], and take a randomly selected sample of 5014 videos, only 22% (1109 videos) contained children performing an action. To measure the effect of training with child-specific data, we use systemic splits in this paper. This allows us to observe visual differences between how and where children perform sporting activities compared to adults.

The main contribution of the paper is in assessing whether there are systemic biases in how a SOTA activity recognition model predicts an activity performed by an adult or a child. To do this, we fine-tuned a SOTA deep learning model on child-specific and adult-specific data separately, and evaluated the resulting models extensively. The following are our contributions:

1) We have collected a child-specific video activity dataset.
2) We performed detailed quantitative and qualitative analyses on the use of SOTA models for child activity recognition.
3) We highlighted the differences between adults and children performing the same actions.

In the next section, we discuss the related work in this area. In Section III, we explain the steps of collecting the datasets in detail. In Section IV, we present the methodology, followed by our experimental results in Section V. In Section VI we conclude this study with a discussion.

[1]This is the uncorrected author proof for Olalere, F., V. Brouwers, M. Doyran, R. Poppe, A.A. Salah, "Video-based sports activity recognition for children," 13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, 2021. Copyright with IEEE: https://ieeexplore.ieee.org/abstract/document/9689651

[2]https://www.play-project.org/

II. Related work

Activity recognition is a problem in computer vision that deals with identifying and classifying different activities in real-life settings from images or video [14]. While there are some standing challenges such as occlusion, differences in how multiple people perform the same action, the application of SOTA deep learning models to current human activity datasets has significantly improved the accuracy at which machines can successfully predict human activities.

One important modeling approach in activity recognition is the multi-stream networks approach [15]. The idea behind multi-stream networks is to model temporal and spatial information using a spatial convolutional neural network (ConvNet) that takes in still images and a temporal ConvNet that takes in motion information from the optical flow field. The output of these ConvNets is fused at some specified convolution layer. Earlier work directly combines the output generated by the softmax layer at the last layer of the network [15]. However, to model spatiotemporal information, earlier interaction between the two streams could also be useful [16], [17].

A more recent network that is modeled after a two-stream network is the SlowFast network [18]. Instead of sampling motion information from the optical flow representation of the video, it simply takes in frames at a different rate to encode spatio-temporal information. The model has achieved SOTA accuracy on datasets such as Kinetics-400 [13], Charades [19], and AVA [20].

Our main focus in this paper is the differences between how actions are performed by adults and children. Children have a different set of body proportions compared to adults, which is important when building automatic analysis tools. Furthermore, children aged between 3-8 years show different motor timing and perception of time than the adults [21]. These differences can be affected by whether or not the child is cooperating with an adult, and even by the cooperation level.

Previous research on child behavior analysis did not necessarily use child-specific training conditions. Deep learning models trained on large datasets, which predominantly included adult data, were used for recognizing gross-motor actions of children [22] or for estimating poses of children during play therapy with an adult [9]. Other researchers created their own network architectures, but still trained them on similar large benchmark datasets with mostly adult videos to recognize actions of children or elderly [23]. To our knowledge, there are no children-specific activity datasets with which child-specific models can be trained.

To test for systemic bias in how a current SOTA deep learning model decodes an activity being done by a child or adult, we collect a sports activity dataset. This domain is a useful starting point, because it is very difficult to collect a new publicly accessible children dataset under strict privacy conditions imposed by regulations and with additional difficulties arising from the Covid pandemic, whereas there are many accessible videos with adequate licensing in the sports domain. Furthermore, the domain is sufficiently challenging, with a lot of activities that are differentiated by subtle visual cues. We discuss the dataset preparation in the next section.

III. Data Collection

To facilitate our research, we collected a sports activity recognition dataset. This dataset is annotated into an adult-specific segment and a child-specific segment. The child-specific segment is made up of videos where children of at most 12 years old perform a specific activity, while the adult-specific segment contains videos of people 13 years and older performing a sporting activity.

After extensive filtering, we selected 21 sport classes from the Kinetics-400 dataset [13] where we could find a sufficient number of videos for both children and adults. Once we selected the classes to work with, we crafted queries to download videos from Youtube with suitable licenses. In crafting the queries, we used targeted words, such as *a child playing basketball, toddlers hitting baseball*. After downloading these videos, we filtered out videos greater than 100MB in size, because we did not want professionally shot videos. The non-professionally shot videos are less edited and are more representative of the real world, in-the-wild setting. Also, this helps reduce the storage requirement and processing time. We also checked that we have a reasonable resolution, which we select to be 720p. Hence, the videos in our dataset can also be used for estimating children's poses.

Next, we split each downloaded video into scenes using *PySceneDetect*[3]. After this, we run a human detector across all the scenes in the video, since we only want clips where a human is performing an activity. To do this, we selected three evenly spaced frames from each scenes that is longer than 1 second and passed these through a pre-trained YOLO-V3 detector [24]. To further filter out videos for the child-specific dataset, we performed automatic child detection. We detect the children using a recently developed zero-shot model called CLIP [25]. We fine-tuned this model with a subset of our dataset following the method proposed by the authors. After trying different configurations, the CLIP configuration that worked best at detecting children in our dataset was using the ImageNet prompt proposed in [25], with each of the sport classes appended to the prompts, a margin scale of 0.2, and zero-padding.

These pre-processing steps left us with just those scenes in each video that had a good chance for the presence of a child, which we estimated as %70. Before we presented the clips for manual annotations, we ran the scenes through a SlowFast model [18] trained on the Kinetics-400 database. We did this to filter out scenes that proved easy to classify, as we wanted to make sure some complexity existed in the dataset. Next, we merged the remaining scenes left after the pre-processing into 60 second long clips, and these clips were passed to manual annotation. For the manual annotation of the clips, we chose to use a crowd-sourcing platform, namely, Amazon's Mechanical

[3]Brandon Castellano, PySceneDetect, available online https://github.com/Breakthrough/PySceneDetect.

Turk (AMT). AMT is often used for tasks such as this [13], [26] and we assumed that since the workers on this platform are more used to such tasks, it increased our chances of getting high-quality annotated data.

For the child-specific dataset, we asked two annotators to select which activity is being performed by a child in the video, we also asked them to indicate at which time the activity starts. Furthermore, we asked them to indicate if the child performs the activity with an adult. This is a useful signal that can be used later for studying interactions between children and adults (see Fig. 1). If there is a disagreement between both annotators, we present the clip to a third annotator. At the end of annotation, we only included clips that were agreed upon by at least two annotators for the child-specific dataset. We additionally annotated 5.014 videos from the Kinetics-400 dataset that fell within the 21 classes we worked with. We asked the annotators to select the videos being performed by children. From this, we got 1.109 videos containing children performing an activity. We downloaded an additional 1.904 videos for the adult-specific dataset to keep the size of both datasets similar.

Fig. 1. Video examples of adults performing activities with a child.

After the annotation, we performed de-duplication to ensure that we only have one clip per Youtube video downloaded in our datasets. We randomly selected one of the annotated clips, if we ended up with more than one clip from the same video in each activity class. While we asked the annotators to indicate if a video contains multiple instances of the activities we are interested in, we only took more than one clip from the video if the video had the word *"compilation"* in its title. A compilation video usually contains clipped together instances of the same or different activities. We had a total of 15 videos with such titles. See Fig. 2 for the download distribution across classes.

Finally, we split the child and adult segments of the dataset into training, validation, and test splits for our experiments. We randomly selected 40 videos per class to be in the training split, which gives 840 videos for the children training data and 840 videos for the adult training data (40×21 classes), respectively. We then split the remaining videos per category in the ratio of 60 to 40 to be in the test and validation datasets. Each of the videos in these datasets are 5 seconds long (See Table I).

IV. Methodology

Using different splits of the data described in Section III, we created five different models to answer our research questions. We start by describing our baseline model, and continue with the Adult Model, the Child Model, the Mixed model (half-split), and the Mixed model (full-split).

	Child-specific dataset	**Adult-specific datatset**
Activity classes	21	21
Total number of videos	1592	1904
Duration per clip	5 seconds	5 seconds
Per second frame rate	30	30
Source	YouTube (483) Kinetics-700 (1109)	Kinetics-700
Total duration	133 minutes	159 minutes

TABLE I
SUMMARY OF THE TWO DATASETS.

A. Baseline Model

A SlowFast model with a ResNet-50 base architecture [18] pre-trained on the HACS-clips dataset [27] serves as the baseline model for all the experiments conducted in this study. The SlowFast model has been shown to achieve state-of-the-art results on activity recognition tasks on benchmark datasets such as AVA v2.1 [20], Charades [28], and the Kinetics-600 dataset [29].

The HACS Clips dataset used for pre-training is a benchmark dataset for activity recognition [27]. It contains 500k annotated video samples spanning across 200 activity classes. The dataset contains videos downloaded from Youtube. After pre-processing, the dataset contains 1.5M 2-seconds clips. The dataset has been shown to outperform other large-scale datasets like Kinetics-600 [29] and Sports1M [30] when used as a pre-training source [27]. We chose to use HACS Clips as our pre-training source because of the similarities it has to our dataset, such as video sources being from Youtube. Also, the taxonomy used in HACS Clips is derived from ActivityNet [26], which shares some taxonomy with Kinetics. We did not use Kinetics as our pre-training source, because we train and test on some videos from the Kinetics-400 dataset.

While HACS Clips bears some similarities with our datasets, issues such as class imbalance within the HACS Clips dataset could lead to a poor generalization of the learned features to our videos. In the Child-specific and Adult-specific datasets, we ensured class balance by selecting the same number of training videos per activity class.

Another source of bias that could arise with this pre-training source is the lack of representation of minority ethnic groups. We see this as a possible source of bias, because the creators of HACS Clips did not explicitly state if measures were taken to prevent such bias [27]. Finally, some of the classes present in our dataset are missing from the HACS Clips dataset and this could affect the generalization of the learned features to our data.

The SlowFast model depicted in Fig. 3 is a a single-stream model that reads in video input at two different frame rates. Unlike two-stream architectures, the SlowFast architecture has a single input source, passed through a Slow pathway and a Fast pathway, which can be any convolutional model that processes videos as a spatiotemporal volume. In our baseline, we use ResNet-50 as our backbone [31]. Both pathways are fused by lateral connections into a SlowFast model [18].

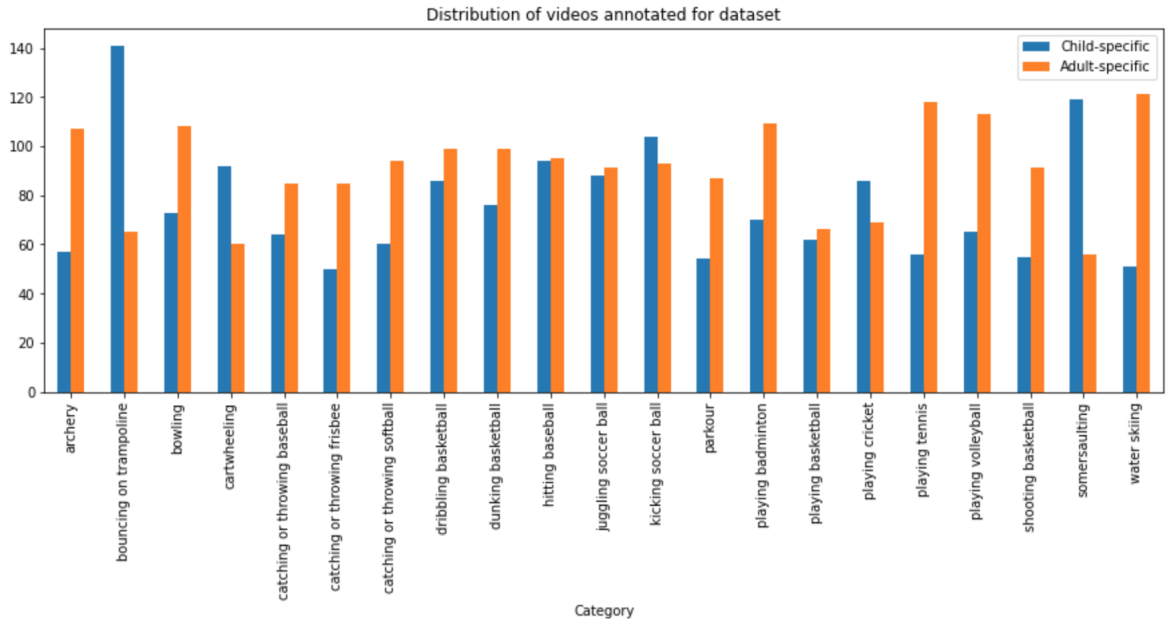

Fig. 2. Distribution of downloaded videos in the Child-specific and Adult-specific datasets.

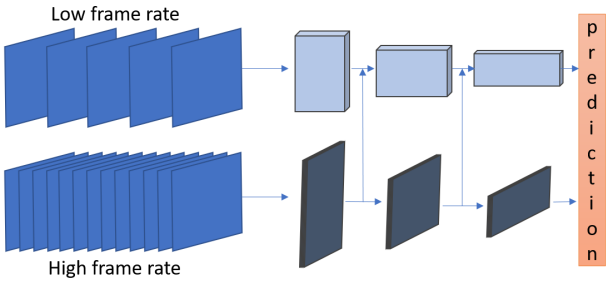

Fig. 3. The SlowFast architecture. Figure adapted from [18].

The idea behind having two pathways is that the categorical semantics of visual contents do not change as rapidly as the motion being performed by the subject. For example, the appearance of a subject performing the activity *catching or throwing baseball* doesn't change much over the frames as with the motion of actually throwing or catching the baseball. Other semantic information like the lighting condition or background colors also would not change so rapidly. Hence, the model can learn spatio-temporal patterns of shorter and longer duration using both pathways.

The Slow pathway requires a low frame rate for processing the input. To do this, we take large strides (16) on our input clips. For the Fast pathway, we take a smaller stride (2). Both pathways are kept informed about what the other one learns through lateral connections after each stage (see Fig. 3). The SlowFast PyTorch implementation we have used is publicly available in a public repository (see footnote). Since this study is not about determining which architecture performs best on our dataset, the SlowFast model is deemed to be sufficient for the experiments we perform. [4]

B. Adult and Child models

We train an Adult model to test the generalizability of features learned from the adult-specific dataset to the child-specific dataset, and a Child model to test the generalization in the other direction.

In both cases, we apply transfer learning to fine-tune the pre-trained model. Given the size of our dataset, we can only fine-tune the last fully connected layer of the pre-trained model without overfitting on the training split. The model contains 34M trainable parameters and we only retrain the last layer with 1.5M parameters, using either the Child-specific or the Adult-specific datasets.

C. Mixed Model (half-split)

As mentioned in Section I, the majority of subjects performing activities in available action recognition dataset are adults. The purpose of creating a Mixed Model is to test if by deliberately ensuring that a dataset has an equal number of children and adult subjects in it, the model can better generalize to both adult-specific and child-specific datasets. The Mixed Model (half-split) will be referred to as the MHS model in this study.

The MHS model is created by fine-tuning the last layer of our pre-trained model using a combination of child and adult-specific datasets. However, we will be using half the training size from each dataset so we end up with the same training size used in both Child and Adult models. This will enable us to see how the training size influences the model's predictions.

[4]https://github.com/sayo20/Video-Based-Sports-Activity-Recognition-for-Children

D. Mixed Model (full-split)

The Mixed model (full-split) is similar to the MHS model. The only difference is that we train it with all the videos available for each class. This means this model is trained with more data than with the previous models described above, which will help illustrating the effect of increasing training set size. We create the model by fine-tuning the last layer of the pre-trained model on both the adult-specific and child-specific training splits. The model will be evaluated against the adult-specific and child-specific test split. The Mixed Model (full-split) model will be referred to as the MFS model in this study.

E. Model Implementation Details

The SlowFast-ResNet50 network is the architecture all our models were built upon. We implemented all our models using Pytorch. We trained the models using an Adam optimizer and used the cross-entropy loss function. We set a batch size of 32 and adjusted the learning rate (starts at 0.01) using the PyTorch *ReduceLROnPlateau*, with a patience value of 80 and a factor of 0.1. The scheduler is called on every batch, and reduces the learning rate by a factor of 0.1 whenever there is no improvement in the loss after 80 batches. To prevent overfitting, we apply early stopping based on the validation loss. If there is no decrease in validation loss after 12 epochs, we stop training the model.

The images are normalized to fit the model input requirements in a way that aspect ratios are retained. All our videos were re-sampled to 30fps to maintain temporal consistency, which ensures that all the video segments cover the same amount of time. We retained the 8fps rate of the pre-training by sampling 40 frames from the 5 second clips. To make sure that the sampled frames spread across a large extent of the clip, we set the stride size as the length of the frame in clips (150) divided by the target number of frames (40). Hence, in this case, we sample every third frame.

We apply the same data augmentation methods discussed in Section III, to keep consistency in the input we present to the models. During training, we apply uniform cropping (256×256) on the videos. Other random augmentations include Gaussian and average blur, left-right flipping, and changes to gamma contrast, linear contrast, hue and saturation. During validation and testing, we only apply a center cropping of 256×256 on the videos.

V. Experimental Results

	Child-specific	Adult-specific
Child Model	43.8%	45.5%
Adult Model	35.0%	51.9%
MHS model	43.8%	51.1%
MFS model	46.2%	52.6%

TABLE II
AVERAGE ACCURACY OF THE MODELS ON CHILD-SPECIFIC AND ADULT-SPECIFIC DATASETS.

In this section, we discuss how each of our models performs on the child-specific and adult-specific test splits of our dataset (see Table II).

Baseline Model: We cannot directly evaluate how the baseline model performs on the child-specific test split because of a mismatch in class labels; our classification problem includes 16 classes that are not included in the baseline model, and only 5 classes are overlapping. We ran some preliminary experiments, and observed that the baseline model performs badly at generalizing what it has learned to the adult-specific and child-specific segments without fine-tuning. Nonetheless, the general features learned from the pre-training source can be useful in fine-tuning the remaining models, given that we have a relatively small dataset.

Child Model: This model achieves an accuracy of 43.8% on the child-specific dataset and 45.5% on the adult-specific dataset for the 21-class problem. We notice that the accuracy is relatively low. Given the properties of the child-specific dataset such as how varied the activities are in terms of how they are performed, where they are being performed, and their challenging camera motions (see Fig. 5), this result is not very surprising. The low accuracy suggests that the videos in the test split are quite different from those in the training split, and the training split might not contain enough variance to generalize to the test splits. Furthermore, while the model performs slightly better on the adult-specific dataset (probably because there is more consistency in the environments and conditions where adult videos are shot), the performance suggests that the Child Model still needs to be trained on more data to do better.

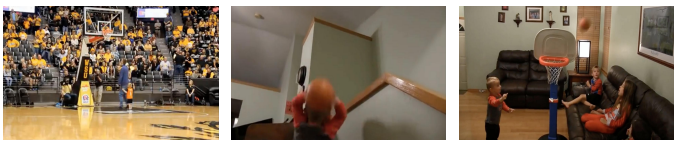

Fig. 4. Children shoot basketball in three different environments.

Fig. 5. Children shoot basketball in three different ways.

By looking at the per-class accuracy on both datasets (see Fig. 8 & Fig. 9), we observe that if the model performs high on a class in the adult-specific dataset, it tends to perform low on the same class within the child-specific dataset, and vice-versa. A general trend we observe with the classes, where the Child model performs badly across both test splits is the variance in the videos within that class. This speaks to the complexity of that class and overall, the dataset. These variations are quite prevalent amongst the videos in the child-specific test split (see Fig. 4), which could also be a reason why the model

performs lower on this split. Furthermore, this model, as the rest of our models, has difficulty in classifying hierarchical classes across both datasets. For example, *playing basketball* is commonly misclassified as *dunking basketball or dribbling basketball* which is a confusion also reported in the Kinetics-400 study [13].

Further studies would be required to test which factors (e.g. video resolution, camera motion, environment, sports equipment, etc.) affect the generalization ability of the children model.

Adult Model: This model achieves an accuracy of 35.0% on the child-specific test split and 51.9% on the adult-specific test split, which clearly illustrates the difficulty of the child-specific dataset. Amongst all the models evaluated on the child-specific dataset (see Table II), the Adult model performs the worst. One reason for this could be that children deviate from standard ways of performing a sports activity in non-trivial ways. Fig. 5 shows for example a way of playing basketball that would not be encountered in an adult-based database. There are some classes within the adult-specific test split that contain more complex videos compared to the child-specific counterpart. For example, the Adult model achieves an accuracy of 62.5% on the archery class in the adult-specific test split and 80.0% on the same class in the child-specific test split. However, we are cautious in interpreting such differences, as the dataset is not too large for sweeping generalizations. A closer look at the archer class reveals that the child-specific test split videos are quite similar to the videos within the adult-specific training split (see Fig. 6 and Fig. 7). We need more test data and experiments to be able to verify exactly why the Adult model generalizes better to some of the classes in the child-specific test split better than the same classes in the adult-specific test split.

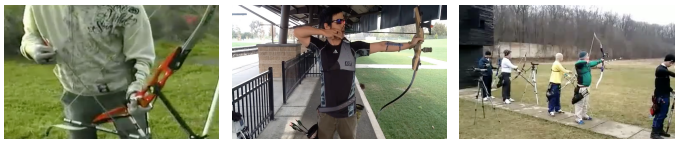

Fig. 6. Sample frames from the archery class in the adult-specific train split.

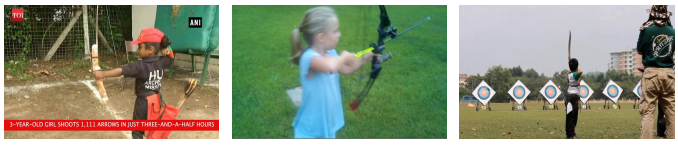

Fig. 7. Sample frames from the archery class in the child-specific test split.

Similar to the Child model, the Adult model also follows the pattern where a higher accuracy of the model on a class within the adult-specific test split is coupled with a lower accuracy of the same class within the child-specific test split, and vice-versa.

We speculate that the model focuses more on environmental features, such as what apparatus is used, and colors in the background, and not on how the activity is being performed, which is why it generalizes badly to the child-specific test split as opposed to the adult-specific test split. This is one important issue in the training of deep learning based models, that they can latch on to a background cue, and derive discriminative information from unexpected features. To overcome this, datasets need to contain large amount of variation. In our case, further studies are required to see which features the models use in making their classifications and how these differ across both datasets.

MHS and MFS Models: The MHS model achieves an accuracy of 43.8% on the child-specific test split and 51.1% on the adult-specific test split. This model was able to match the performance of the Child model on the child-specific test split (43.8%) and performs a little bit lower than the Adult model on the adult-specific test split (51.9%). However, the MHS model only uses half the training size compared to both the Child and Adult models.

When we increase the training size of the MHS model and switch to the MFS model, we see an increase in performance on both the adult-specific and child-specific test splits, as expected (see Table II). This suggests that by including both child-specific and adult-specific videos for training, we increase the overall generalizability of the model to both adult-specific and child-specific datasets.

However, when we compare how the MHS model performs on the adult-specific test split as opposed to how the Adult model performs on the same split, the MHS model does not seem to benefit a lot from including children videos in the training sample, in comparison to the increase in performance we see with the MHS model on the child-specific test split. We would require an ablation study to determine the effect of including both data types for training and what balance of child-specific and adult-specific data would be required to increase the model's generalization to both data types. A similar setup is required to determine whether including both types is really beneficial for generalization, or if just increasing the training size of both Adult and Child models proves best for the generalizability of the models for the corresponding sets.

VI. Conclusion

This study investigated whether SOTA deep learning models for action recognition trained on mostly adult videos can generalize well on a child-specific dataset. To do this, we created a child-specific and an adult-specific dataset, as well as models based on different types of datasets. We used a single deep learning architecture, but we believe our results will extend to other architectures under similar conditions. We presented our data collection pipeline in detail, to foster future collection of children-specific data.

Our results show that, while SOTA deep learning can be used to classify children's sporting activities, this is more difficult compared to adult sporting activities, because children display higher variance in this domain. In other activities and interactions, this may not be the case, as adults can engage

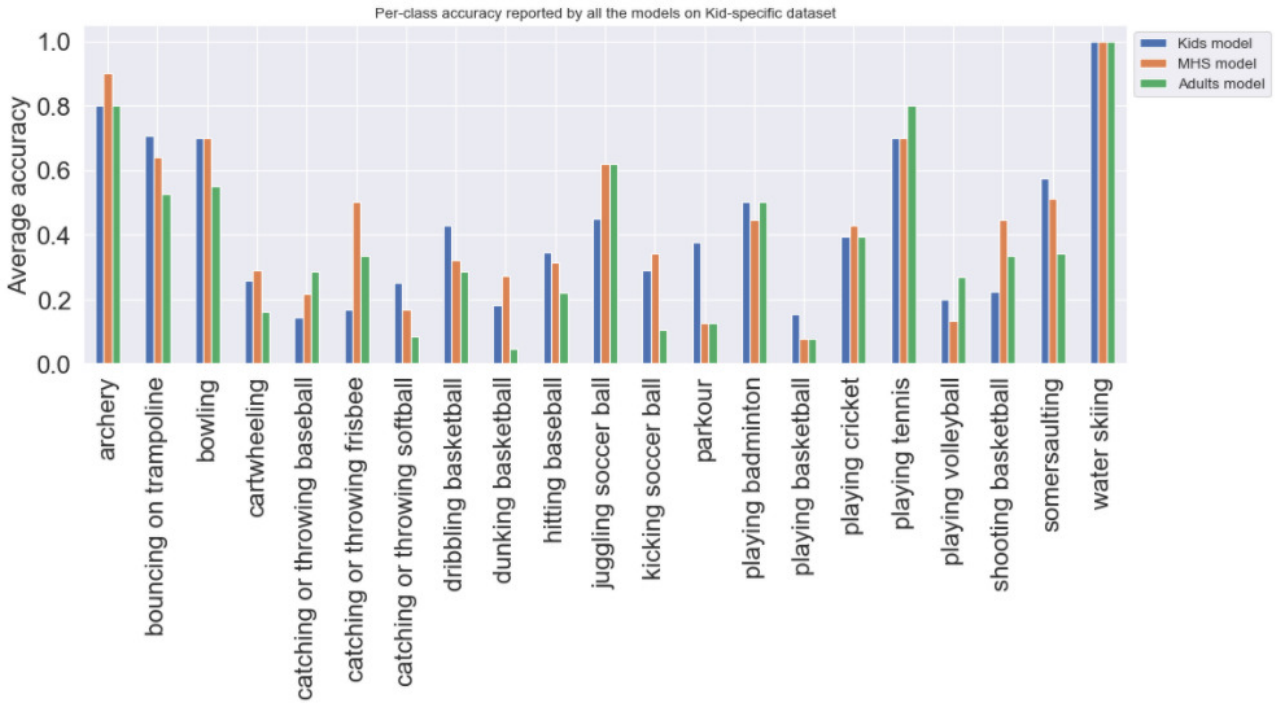

Fig. 8. Per-class accuracy reported by the Child model, MHS model, and Adults model on the child-specific test split.

in many more activities compared to children. For many sports, however, adult behavior is more regular, and performed in specific settings, which provides consistency and reduced variance in the background as well.

Our study also shows that the features learned from training on a child-specific sports activity dataset alone can be used to classify adult sports activities, while the reverse is not the case. Also, including both adult and children videos in training improves generalization. More work is needed to determine the effect of fine-tuning in other activity recognition domains, and in specific settings like at home play settings, and parent-child interactions, which are important application areas of child behavior analysis.

REFERENCES

[1] Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[2] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology*, 23(11):1993–2008, 2013.

[3] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[4] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[5] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proc. ECCV*, pages 0–0, 2018.

[6] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R Pierce, Daniel K Bogen, Laura Prosser, Michelle J Johnson, and Konrad P Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020.

[7] N Charlotte Onland-Moret, Jacobine E Buizer-Voskamp, Maria EWA Albers, Rachel M Brouwer, Elizabeth EL Buimer, Roy S Hessels, Roel de Heus, Jorg Huijding, Caroline MM Junge, René CW Mandl, et al. The youth study: Rationale, design, and study procedures. *Developmental cognitive neuroscience*, 46:100868, 2020.

[8] Ori Ossmy, Rick O Gilmore, and Karen E Adolph. Autovidev: A computer-vision framework to enhance and accelerate research in human development. In *Science and Information Conference*, pages 147–156. Springer, 2019.

[9] Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, Sibel Halfon, and Albert Ali Salah. Video and text-based affect analysis of children in play therapy. In *2019 International Conference on Multimodal Interaction*, pages 26–34, 2019.

[10] Sibel Halfon, Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, and Ali Albert Salah. Multimodal affect analysis of psychodynamic play therapy. *Psychotherapy Research*, 31(3):313–328, 2021.

[11] Metehan Doyran, Arjan Schimmel, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander CJ Bakkes, Heysem Kaya, Ronald Poppe, and Albert Ali Salah. Mumbai: multi-person, multi-modal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces*, pages 1–19, 2021.

[12] Albert Ali Salah, Jeffrey Cohn, Ronald Poppe, and Heysem Kaya. Computational and affective approaches to behavioral and clinical science. *Submitted for publication*, 2021.

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[14] Eunju Kim, Sumi Helal, and Diane Cook. Human activity recognition and pattern discovery. *IEEE pervasive computing*, 9(1):48–53, 2009.

[15] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. CVPR*, pages 1933–1941, 2016.

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

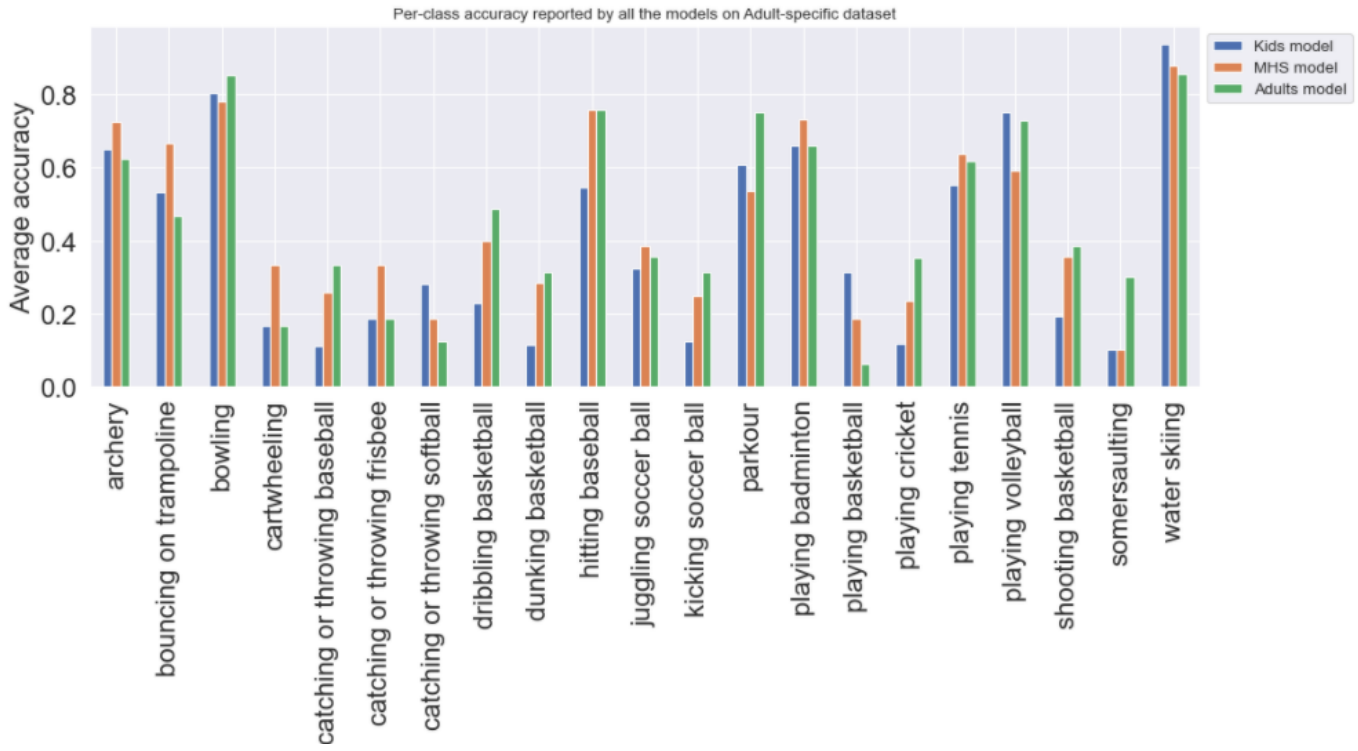

Fig. 9. Per-class accuracy reported by the Child model, MHS model, and Adults model on the adult-specific test split.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. ICCV*, pages 6202–6211, 2019.

[19] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[20] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

[21] Florie Monier and Sylvie Droit-Volet. Synchrony and emotion in children and adults. *International Journal of Psychology*, 53(3):184–193, 2018.

[22] Satoshi Suzuki, Yukie Amemiya, and Maiko Sato. Enhancement of gross-motor action recognition for children by cnn with openpose. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, volume 1, pages 5382–5387. IEEE, 2019.

[23] Chang-Di Huang, Chien-Yao Wang, and Jia-Ching Wang. Human action recognition system for elderly and children care using three stream convnet. In *2015 International Conference on Orange Technologies (ICOT)*, pages 5–9. IEEE, 2015.

[24] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[26] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proc. CVPR*, pages 961–970, 2015.

[27] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proc. ICCV*, pages 8668–8678, 2019.

[28] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[29] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, pages 1725–1732, 2014.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.