A Detection-based Approach to Multiview Action Classification in Infants

Carolina Pacheco*†, Effrosyni Mavroudi*†, Elena Kokkoni‡, Herbert G. Tanner§ and René Vidal*†
* Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 21218, USA.
† Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA.
‡ Department of Bioengineering, University of California, Riverside, CA 92521, USA.
§ Department of Mechanical Engineering, University of Newark, Delaware, DE 19716, USA.
{cpachec2, emavrou1, rvidal}@jhu.edu, elena.kokkoni@ucr.edu, btanner@udel.edu.

***Abstract*—Activity recognition in children and infants is important in applications such as safety monitoring, behavior assessment, and child-robot interaction, among others. However, it differs from activity recognition in adults not only because body poses and proportions are different, but also because of the way in which actions are performed. This paper addresses the problem of infant action classification in challenging conditions. The actions are performed in a pediatric rehabilitation environment in which not only infants but also robots and adults are present, with the infant being one of the smallest actors in the scene. We propose a multiview action classification system based on Faster R-CNN and LSTM networks, which fuses information from different views by using learnable fusion coefficients derived from detection confidence scores. The proposed system is view-independent, learns features that are close to view-invariant, and can handle new or missing views at test time. Our approach outperforms the state-of-the-art baseline model for a small dataset (2 subjects, 10-24 months old) by 11.4% in terms of average classification accuracy in four classes (crawl, sit, stand and walk). Moreover, experiments in an extended dataset (6 subjects, 8-24 months old) show that the proposed fusion strategy outperforms all the alternative fusion methods studied.**

I. Introduction

Human activity recognition has drawn the attention of computer vision researchers for many years, due to its applications in fields as diverse as human-computer interaction [1], surveillance [2] and health care [3]. The main goal of human activity recognition is to identify actions performed by one or more humans in a temporal sequence of observations. The increased computational power together with the availability of public datasets with several thousands of annotated clips have allowed for the training of large networks, such as multi-stream 3D CNN architectures [4], significantly boosting action recognition performance. However, human activity recognition research has so far been primarily focused on adults.

Children activity recognition has important applications to safety monitoring [5], object-play behavior assessment [6], real-time feedback provision to adaptive environments [7], [8], [9], [10], and others. However, children are largely underrepresented in both human pose estimation datasets [11] and human activity recognition datasets. This can be explained by the fact that data from adult activities can be used in many more applications than data from children activities. Moreover, some of the largest annotated human activity datasets rely on user-uploaded videos to public platforms [12], [13], [14], in which the presence of children is limited due to privacy concerns.

Challenges. Activity recognition in children differs from activity recognition in adults not only because of different body poses and proportions [11], but also because of the way in which actions are performed. In fact, a study of 5 to 8 year-old children showed that classifying actions is harder in children than in adults, and that adult-trained systems do not transfer well to children [15]. The latter was also recently confirmed in [7] by a large-scale study in children aged 6 to 10 years. Thus, given its discrepancies with standard adult activity recognition, children activity recognition requires special attention.

In the context of early pediatric rehabilitation, Kokkoni et al. [10] recently proposed a learning environment for infants in which interaction with socially assistive robots and body weight support technology are used to promote the infants' mobility. Sessions are recorded from a set of cameras, and the system can benefit from infant activity recognition for both real-time feedback to the robots and assessment of mobility outcomes. However, this task is challenging because: (i) the poses observed in infants are significantly different from those observed in adults, and therefore off-the-shelf pre-trained activity recognition systems would not work, (ii) infants are often occluded by other actors or elements in the scene (such as adults, robots, objects in the environment, etc), and thus the information from a given camera is not always useful, and (iii) the body size of infants is not only smaller than the size of other actors, but also significantly smaller than the frame size (covering in average less than 1.7% of its area).

Paper contributions. In this paper we propose a multiview action classification system for infants that first localizes the subject in each view, extracts spatio-temporal features from a tubelet around the subject, fuses these features across views, and then classifies the action. More specifically, spatial localization is performed by a view-independent detector based on Faster R-CNN [16] fine-tuned to detect infants. The features obtained by the detector are subsequently fed to a view-independent temporal model (a recurrent neural network (RNN) with long short term memory (LSTM) units [17]) to obtain a spatio-temporal representation of the tubelet around the subject. Fusion coefficients based on detection confidence

scores are then learned for combining information across different views. Experiments performed in multiview data from 6 subjects with an average age of 12.7 months show that our approach not only significantly outperforms the state-of-the-art baseline model for this dataset [10] (11.4% improvement in average classification accuracy in four classes), but also outperforms all other fusion strategies studied in this paper. In summary, the contributions of this paper are:

1) A detection-based multiview action classification system specifically designed for scenes with a single (small) subject of interest moving in complex scenes.
2) A novel and meaningful use of the detection confidence scores to learn fusion coefficients that adaptively combine information from different views.
3) State-of-the-art action classification results for rehabilitation therapy in infants, which show that our system performs well in terms of generalization to new subjects.

The proposed system directly addresses the challenges imposed by the application. The infant detector and action classifier are trained with data from infants, which allows learning relevant patterns specific to the population of interest. Moreover, the use of an infant detector not only helps to overcome the small relative size of the infants in the scene, but also provides information related to their visibility, which is used to adaptively weigh the importance of each view.

While temporal action localization is out of the scope of this paper, the proposed method can be easily extended to perform action segmentation given the flexibility of RNNs.

II. Related work

Given the vast literature in activity recognition and multiview learning in general, this section focuses on children activity recognition in particular. More general approaches are discussed if they relate directly to the proposed method.

Early work on children activity recognition was based on sensors such as accelerometers and barometers attached to wearable devices [5], [15], [18]. Most of these papers used traditional machine learning methods such as Support Vector Machines (SVM) or Self Organizing Maps (SOM) to classify actions based on handcrafted features, although recent work has leveraged deep models [19], [20]. Wearable devices are useful to capture data directly from the subject of interest, but they can interfere with the typical behavior of the child, especially in the early age [21]. Therefore, in this work we focus on computer vision-based activity recognition.

A line of research of growing interest in computer vision-based children activity recognition is motion analysis of newborns and young infants (0-6 months old), given its relevance in monitoring respiratory events [22] and early detection of movement disorders [23], [24]. They leverage the limited range of movement of such young infants and locate a camera above and in parallel to the crib surface, to acquire videos with the infant as the only actor covering a significant portion of the image. These conditions do not apply when studying older infants in more complex environments.

More challenging scenarios have been studied in the field of child-robot interaction, where computer vision-based children activity recognition has been performed in multiview setups [7], [8], [10]. In [7] the use of deep versus handcrafted dense trajectories (DT) features [25], as well as fusion of information from different views is studied in SVM classifiers. It was found that early fusion at the feature level is more effective for DT features in bag of visual words encodings (DT-BOVW), while late fusion at the score level works better for deep features. In [8] DT-BOVW and SVM were used, and different post-processing fusion strategies at the score level were studied to perform action classification. It was concluded that maximizing the mean score or the minimum score of the selected class works better than maximizing the maximum score as a post-processing fusion strategy. While these models are general, they were tested on 6 to 10 year-old children whose actions are more similar to those of adults than to those of infants. In contrast, DT-BOVW were used to perform action classification in 10 to 24 month-old infants in [10], where the multiview problem was addressed by a multiple instance learning SVM scheme (MI-SVM), considering views as instances of the same sample and taking into account that the action might not be observed in all of them. That strategy outperformed post-processing fusion strategies, however its classification performance was modest; arguably, it did not address the challenges imposed by the complexity of the scene.

Here, we directly address these challenges by using local features from spatial regions of interest—a common practice in activity recognition. Spatial localization of relevant regions is usually achieved by motion estimation [25] or person detection [8]. Faster R-CNN [16] is a well-known object detector that has been trained in large-scale image datasets. Naturally, there have been efforts to use this architecture for spatio-temporal localization [26], and to extend it to generate spatio-temporal tubelets of interest [27]. In our case, the availability of annotations from isolated non-consecutive frames allows for training and evaluation of frame-by-frame detections, but not for spatio-temporal localization. Nonetheless, this still succeeds in extracting the local information that is relevant for the action classification task. By leveraging the object detector not only by feeding the extracted features to the classifier, but also by using the detection confidence scores to fuse the information from different views in an adaptive way, we outperformed all other fusion strategies we explored.

One of the main advantages of the availability of data from different views is the variability of information they can capture. However, this variability is also the main challenge. Generally speaking, the action to classify is the same regardless of the viewpoint. Therefore, the key to address this problem is to seek for view-invariant features. One way to do so is to jointly use multiple views to infer 3D information, and then extract view-invariant features from it [9]. Alternative approaches include cross-view learning (where features from a given view are estimated based on observations from another view [28]), directly learning to map any view to a canonical view [29], or learning a new shared representation based

on similarities among different views [30]. Inference of 3D information, cross-view learning and canonical view mapping either require large amounts of data, which comes with significant computational cost, or a good initialization point. As discussed above, action recognition in infants is significantly different from traditional action recognition, which severely limits the availability of good models that can be used for initialization purposes. Therefore, what we propose instead is the use of a view-independent architecture that generates close to view-invariant spatio-temporal features, given its shared architecture and joint supervision for the multiple views. In that sense, our approach is akin to learning a new shared representation. Experiments show that these close-to-view-invariant features learned can improve performance even when data from only one view are used at test time. The use of a view-independent architecture also has practical advantages, such as simultaneously learning from more and diverse data.

III. Methods

Let $\{X_n\}_{n=1}^N$ be a set of synchronized $\mathcal{T}$ frame-long video clips captured from N different viewpoints of the same scene. Each one of the clips can be written as $X_n = \{x_n^1, \ldots, x_n^{\mathcal{T}}\}$ for $n = 1, \ldots, N$, where x_n^t represents the t^{th} frame of the n^{th} view. Assume the scene contains only one actor of interest (infant) in each clip and that the actor performs a single action $y \in \{0, 1, \ldots, C-1\}$, where C is the total number of classes.

In this section, we consider the problem of assigning a label $\hat{y} \in \{0, 1, ..., C-1\}$ to a video clip $\{X_n\}_{n=1}^N$ such that $\hat{y} = y$, i.e., the action is correctly classified. We propose a view-independent detection-based multiview action classification system, whose overview is depicted in Fig. 1 for the case $N = 3$. The main components of this system are an infant detector, an LSTM encoder, and learnable fusion coefficients based on detection confidence scores. The components are view-independent, and they are described in detail next.

A. Infant detector

In principle, any off-the-shelf person or infant detector could be used to localize the infant in the frames; however, most state-of-the-art object detectors have been trained on datasets in which the "infant" or "child" categories do not exist and the "person" category has a significant bias towards adult subjects. Therefore they fail to recognize infants, especially in challenging scenes. We decided to fine-tune the state-of-the-art object detector Faster R-CNN [16] for this purpose.

The Faster R-CNN architecture is composed of a region proposal network (RPN) in which several regions of the image are refined and evaluated in terms of their probability of containing an object (*objectness*), and an object classifier in which the regions with highest *objectness* are further refined and evaluated in terms of their probability of belonging to each one of the classes of interest. The RPN and the object classifier share a convolutional neural network (CNN) as a feature extractor, making the overall detection process computationally efficient. Our infant detector uses a Faster R-CNN architecture with ResNet-101 as feature extractor, which had been pre-trained in the large-scale object detection dataset COCO. We modified the last layer of the classifier to output the probability that the object belongs to just one class (instead of the 80 classes in COCO), and then trained the object detector with examples of bounding boxes containing infants performing actions in a rehabilitation therapy setting. See section IV-C for details on the implementation and training.

At inference time, each frame x_n^t for $n = 1, \ldots, N$ and $t = 1, \ldots, \mathcal{T}$ is processed by the object detector which outputs a set of K candidate infant detections, where each detection is described by a bounding box $b_{n,k}^t \in [0,1]^4$ and a corresponding confidence score $s_{n,k}^t \in [0,1]$ for $k = 1, \ldots, K$. Detection is performed frame by frame, so we apply a smoothing strategy to leverage information from neighboring frames and provide smoother detections to the action classifier. The average location of the center of the infant is computed in a sliding window fashion: considering only the most confident detection per frame, the one-third frames with highest confidence within the window are used to compute the average location. For each frame we then select the most confident detection unless its distance to the average center location is large. If that is the case, we select the closest detection to the average center instead.

Let κ_n^t be the index of the candidate detection selected for the t^{th} frame of the n^{th} view. Then, for each $\mathcal{T}$-long clip X_n, the infant detector outputs a sequence with the selected locations of the infant $B_n = \left[b_{n,\kappa_n^1}^1, \ldots, b_{n,\kappa_n^{\mathcal{T}}}^{\mathcal{T}}\right] \in \mathbb{R}^{\mathcal{T}\times 4}$ and their corresponding detection confidence scores $S_n = \left[s_{n,\kappa_n^1}^1, \ldots, s_{n,\kappa_n^{\mathcal{T}}}^{\mathcal{T}}\right] \in \mathbb{R}^{\mathcal{T}}$. We use feature maps already computed for detection (from ResNet-101 in this case) to provide local features from the selected locations as inputs to the downstream task

$$\mathcal{F}(X_n, B_n) = \left[\Phi\left(x_n^1, b_{n,\kappa_n^1}^1\right), \ldots, \Phi\left(x_n^{\mathcal{T}}, b_{n,\kappa_n^{\mathcal{T}}}^{\mathcal{T}}\right)\right] \quad , \quad (1)$$

where the first argument of the feature map $\Phi(\cdot)$ indicates the whole frame used as an input, and the second argument represents a region to be cropped from the feature map.

B. LSTM encoder

Long Short Term Memory units (LSTMs) have been widely used in many fields, including activity recognition [9], [19], [31]. Their learnable gated connections were proposed to overcome the vanishing gradient problem observed in training recurrent neural networks [17], and nowadays LSTMs are one of the standard architectures used to process time series.

The LSTM network serves as an encoder $\mathcal{G}(\cdot)$ that maps the time series $\mathcal{F}(X_n, B_n)$ in (1) to a vector f_n as

$$f_n = \mathcal{G}\big(\mathcal{F}(X_n, B_n)\big) \in \mathbb{R}^H, \quad n = 1, \ldots, N \quad , \quad (2)$$

where H is the dimension of the hidden layer of the LSTM. As is a common practice in action classification approaches, f_n simply corresponds to the final LSTM state. We then apply a linear layer to map the clip encodings to the number of classes

$$\tilde{y}_n = A^T f_n + a \in \mathbb{R}^C, \quad n = 1, \ldots, N \quad , \quad (3)$$

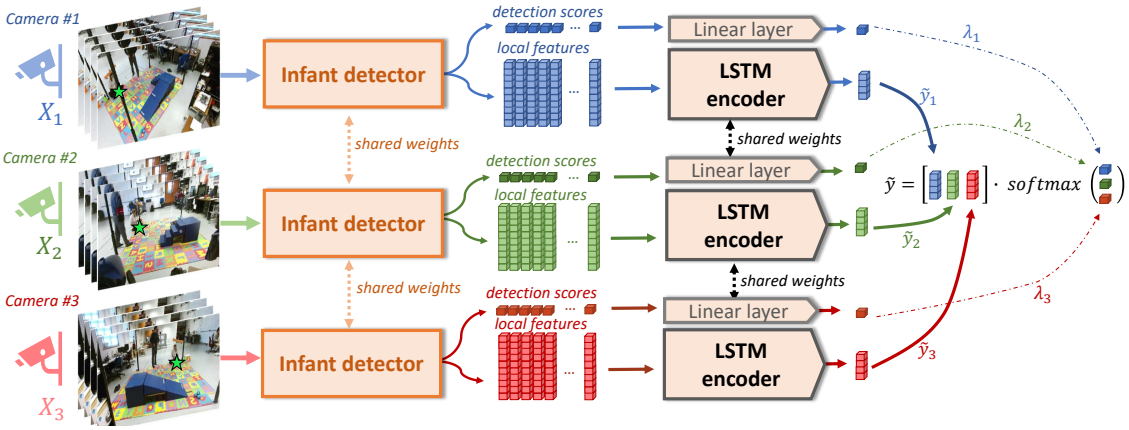


Fig. 1: Overview of the detection-based multiview action classification system for $N = 3$ views. The system first localizes the infant in the different views and then it adaptively combines their local information to classify the actions. Green stars ★ mark the location of the infant in the input images as a reference for the reader.

where $A \in \mathbb{R}^{H\times C}$ and $a \in \mathbb{R}^C$ are learnable weights. The quantities $\tilde{y}_n$ are referred to as *logits*. The LSTM encoder can be trained to perform single-view action classification by minimizing the cross-entropy loss $\ell(y, \tilde{y}_n)$ between the ground truth class y and $\tilde{y}_n$ for $n = 1, \ldots, N$ separately. However, in the proposed view-independent architecture, $\mathcal{G}(\cdot)$, A and a are shared across views—see Section III-C.

C. Learnable fusion coefficients

To combine the information from different views, we propose to learn fusion coefficients $\lambda_n = P^T S_n + p \in \mathbb{R}, n = 1, ..., N$, where $P \in \mathbb{R}^{\mathcal{T}}$ and $p \in \mathbb{R}$ are learnable weights used to combine the detection confidence scores S_n from different frames of the same view. The fusion coefficients are then normalized to ensure they are non-negative and add up to one:

$$\tilde{\lambda}_n = \frac{e^{\lambda_n}}{\sum_{i=1}^{N} e^{\lambda_i}} = \frac{e^{P^T S_n + p}}{\sum_{i=1}^{N} e^{P^T S_i + p}} \in [0, 1], \quad n = 1, \ldots, N. \quad (4)$$

The *fused logits* $\tilde{y}$ are then obtained by a linear combination of the logits

$$\tilde{y} = \sum_{n=1}^{N} \tilde{\lambda}_n \tilde{y}_n \in \mathbb{R}^C \ . \quad (5)$$

The introduction of learnable fusion coefficients $\tilde{\lambda}_n$ based on detection confidence scores allows the classifier to adaptively weigh the contribution of different views, making the action classifier rely more on views in which the detector is more confident or the infant is more visible, as opposed to views in which the infant might be wrongly detected or occluded.

The parameters of the LSTM encoder and the parameters of the fusion coefficients are then jointly learned by minimizing

$$\mathcal{L} = \ell(y, \tilde{y}) + \sum_{n=1}^{N} \ell(y, \tilde{y}_n) \ , \quad (6)$$

where $\ell(\cdot, \cdot)$ corresponds to the cross-entropy loss, y is the ground truth action class, $\tilde{y}$ corresponds to the fused logits from the proposed multiview system, and $\tilde{y}_n$ correspond to the logits predicted independently for each one of the N views.

At inference time, the classification *score* of each class is obtained by applying the softmax function to $\tilde{y}$, and the predicted class $\hat{y}$ for the set $\{X_n\}_{n=1}^N$ is the one that maximizes the classification score. In summary, the proposed architecture and fusion strategy generate a view-independent system whose output is invariant to permutations of the views. Moreover, it is flexible enough to handle samples with different number and combination of views.

IV. Experiments

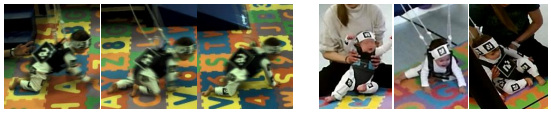
A. Data acquisition and annotation protocol

Six non-walking infants between the age of 7.8 and 23.7 months at study onset ($\mu = 12.7$, $\sigma = 5.6$), one of them diagnosed with Down Syndrome (DS), participated in 8 one-hour sessions over the course of 4 weeks. At each session, infants engaged in motor tasks while interacting with two socially assistive robots. In half of the trials, infants used the assistance from a body weight support system, which provides mechanical support to ease mobility in an open area [32], [33], [34]. Infants wore a white suit with printed AR tags on certain parts of the body. The sessions were recorded through a network of five cameras (KINECT, from Microsoft). Out of all the different motor tasks infants participated in, here we focus on the crawling, walking, and standing tasks. In the crawling and walking tasks, infants were encouraged to move using the corresponding form of locomotor action while following the robots (infants were assisted by the researcher in the walking trials). In the standing task, infants were engaged in manipulation of a chest-high multi-sensory table toy with the humanoid robot. The researcher was always in the scene and close to the infant. As is typical in natural play-type scenarios with developing infants, the trials involved more than the desired infant motor action; in fact, varying (unplanned) spontaneous motor actions were demonstrated.

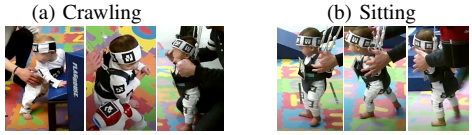
For localization annotation, the coordinates of a bounding box containing the infant's body were annotated using open-

(a) High visibility (b) Medium visibility (c) Low visibility

Fig. 2: Examples of bounding box annotations for infant detection. Colors of the bounding boxes indicate different visibility levels.



(a) Crawling (b) Sitting



(c) Standing (d) Walking

Fig. 3: Examples of cropped frames from action annotated clips.

source software (www.kinovea.org [v. 0.8.27]). A new bounding box was annotated every time the infant changed motor action or moved by more than 50% out of the frame of the previous bounding box. In cases where the infant had remained stationary for a long time, a bounding box was annotated in between. At most two views per clip were annotated. Three infant body visibility levels were defined for bounding boxes: high (more than 80% of the body is visible), medium (between 20% and 80% of the body is visible), and low (less than 20% of the body is visible). Examples of bounding box annotations can be found in Fig. 2.

For action annotation, the main motor actions seen in infant development were annotated using a Java-based application widely used in developmental research (www.datavyu.org, [v.1.5.0]). The following locomotor and postural actions were annotated: crawling, walking, sitting, and standing. In cases infants demonstrated other actions, these were not annotated. Examples of frames from annotated clips are shown in Fig. 3.

B. Cross-validation strategies

In total, 13.839 bounding boxes from five cameras corresponding to high (83.13%), medium (14.46%) and low visibility (2.41%), and 1.603 instances corresponding to crawl (17.97%), sit (24.27%), walk (36.62%) and stand (21.15%) were annotated. Experiments were performed following two cross-validation strategies: (i) **Leave One Super-session Out**, in which 2 out of the 8 sessions from each one of the subjects were held out for testing, and (ii) **Leave One Subject Out**, in which data from one of the subjects were held out for testing. Unfortunately, in the first 7 out of the 8 sessions, subject T1 did not wear the same type of tagged white suit that the other subjects did, which introduced high unexpected variability when compared to the rest of the subjects. Therefore, for the sake of comparability, we considered only the last session from this subject, and combined it with data from the subject that had fewer bounding box annotations (T5) for cross-validation. Annotations available for each split are detailed in Table I.

Splits	**Boxes**	**Actions**				
		Crawl	Sit	Stand	Walk	**Total**
Super-session 1	3420	73	129	133	85	420
Super-session 2	3211	63	86	142	88	379
Super-session 3	3525	73	79	156	91	399
Super-session 4	3683	79	95	156	75	405
Total	**13839**	288	389	587	339	**1603**
Subject T2	5171	107	154	322	157	740
Subject T3	0	30	86	44	24	184
Subject T4	3113	58	80	69	64	271
Subject T5 & T1	2723	57	37	90	57	241
Subject T6	2832	36	32	62	37	167
Total	**13839**	288	389	587	339	**1603**

TABLE I: Annotations available for test splits in Leave One Super-session out (top) and Leave One Subject Out (bottom) cross-validation strategies. Training data for each split correspond to the sum of the data in the remaining splits.

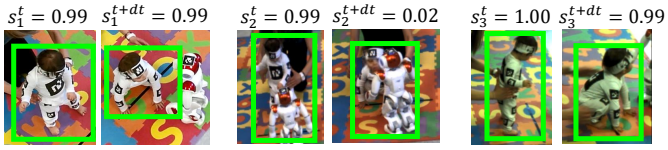
Splits	**Boxes**	**Actions**				
		Crawl	Sit	Stand	Walk	**Total**
Split 1	1977	34	47	123	64	268
Split 2	1849	31	51	89	65	236
Split 3	1969	47	70	126	48	291
Total	**5795**	112	168	338	177	**795**
Additional data	**2366**	54	60	68	71	**253**

TABLE II: Annotations available for test splits from data studied in [10]. Training data for each split correspond to the sum of the data in the remaining splits plus the additional data.

To compare our method with the one presented in [10], we performed experiments considering only data from the two subjects studied in [10] (T1 and T2). Instead of randomly splitting the data as done in [10], we performed cross-validation in a Leave One Super-session Out fashion. Since one of the subjects corresponds to the infant discussed above, only 9 sessions were used for test purposes, divided in 3 splits. The amount of training data in this case is significantly less, and therefore we included sessions 1 to 7 from T1 as additional training data for all splits in both methods; see Table II.

C. Implementation details

To detect infants, the Tensorflow Object Detection API [35] was used to fine-tune a Faster R-CNN architecture pre-trained on COCO, with ResNet-101 as feature extractor. The first three blocks of the ResNet-101 architecture were frozen to reduce the number of trainable parameters and prevent overfitting. Vertical and horizontal flipping as well as jittering were used for data augmentation. The maximum number of proposals was set to $K = 300$, and the remaining parameters were set as default. The detector was fine-tuned for 150.000 iterations. Regardless of the action classification scheme, the detector was always trained using data from five different views. For temporal smoothing of the detections, a sliding window of 30 frames and stride 5 with a threshold of $\sqrt{0.0005}$ normalized pixels provided desirable qualitatively results in the videos. Features from the third block of the ResNet-101 were extracted in the detected locations and reshaped to $14 \times 14 \times 1024$-sized feature maps per frame. Spatial average pooling of the top



(a) Camera #1 (b) Camera #2 (c) Camera #3

Fig. 4: Examples of detections in synchronized data from different cameras. s_n^t corresponds to the detection confidence score in frame t, view n. Low score is observed when the infant is occluded.

half (first 7 rows) and bottom half (last 7 rows) of the feature maps was then performed to preserve some spatial information and end up with a 2048-sized vector feature per frame. For the LSTM encoder, a two-layered network with hidden size $H = 512$ was implemented in Pytorch. Number of classes was set to $C = 4$ corresponding to crawl, sit, stand and walk. Oversampling was performed to balance the number of instances per class in the training set, and maximum number of frames was set to $\mathcal{T} = 150$. Crossentropy loss and L_2 regularization with a factor of 5×10^{-5} were used to train the classifiers for 1.000 iterations at an initial learning rate of 3×10^{-5} with RMSProp optimizer and a dropout [36] rate of 0.5. Batch size was set to 64 samples.

D. Detection results

Table III presents detection results for both cross-validation strategies. The infant is considered successfully detected if the intersection over union between the most confident detection for the frame and the annotated box is larger than 0.5. On average, the infant detector achieves 95.2% and 94.6% accuracy in the Leave One Super-session Out and Leave One Subject Out experiments, respectively. Moreover, in both cases the detector accuracy is directly related to the level of visibility of the infant, achieving less than 37% accuracy on average in cases of low visibility, and around 98% in cases where the infant visibility is high. Examples of detections are depicted in Fig. 4. This evaluation of the detection results is pessimistic for the purpose of the action classifier because (i) the evaluation is performed frame-by-frame based on the most confident detection only, as opposed to our system that temporally smooths the detections, and (ii) the detection evaluation considers only one view at a time, while our system uses data from different views (in which at each time the infant is likely visible from at least one of them). Hence, given the lack of annotations to evaluate multiview detection, these results are considered sufficient as a first step to our system.

E. Baseline action classification methods

Single view. It corresponds to the case in which data from one camera are used to train a view-specific LSTM encoder and the classification is based on the logits from that given view.

Post-processing fusion. In this case, view-specific LSTM encoders are trained independently, and information from different views is combined as a post-processing step. Specifically, let $\tilde{y}_n$ be the logits from the view-specific classifier n, and $\sigma(\tilde{y}_n)$ the corresponding classification score of each class, where $\sigma(\cdot)$ denotes the softmax function. The action prediction $\hat{y}$ of different post-processing strategies is described below.

max score:	$\underset{c\in[0,\dots,C-1]}{\text{argmax}}$	$\max_{n\in[1,\dots,N]} \sigma(\tilde{y}_n)_c$
mean score:	$\underset{c\in[0,\dots,C-1]}{\text{argmax}}$	$\frac{1}{N}\sum_{n=1}^{N} \sigma(\tilde{y}_n)_c$
sum logits:	$\underset{c\in[0,\dots,C-1]}{\text{argmax}}$	$\sum_{n=1}^{N} \tilde{y}_{n,c}$

Joint training fusion. It corresponds to training a view-independent architecture. The **sum logits** case refers to the last method described in the post-processing fusion section, except that the LSTM encoder is jointly trained for all views.

Splits	**Visibility**			
	Low	Medium	High	Total
Super-session 1	40.5	89.5	98.8	96.3
Super-session 2	28.9	87.1	98.3	95.2
Super-session 3	32.9	86.9	97.8	94.6
Super-session 4	22.0	85.9	98.1	94.6
Average	**31.1**	**87.4**	**98.2**	**95.2**
Subject T2	27.9	78.8	96.7	91.1
Subject T3	—	—	—	—
Subject T4	34.6	81.9	98.3	95.3
Subject T5 & T1	37.8	83.3	97.4	94.3
Subject T6	47.4	91.3	99.3	97.6
Average	**36.9**	**83.8**	**97.9**	**94.6**

TABLE III: Detection accuracy (%). The infant is considered detected if the most confident prediction intersects the annotated bounding box with intersection over union larger than 0.5. "Total" is not the average over the other columns because data per visibility are not balanced.

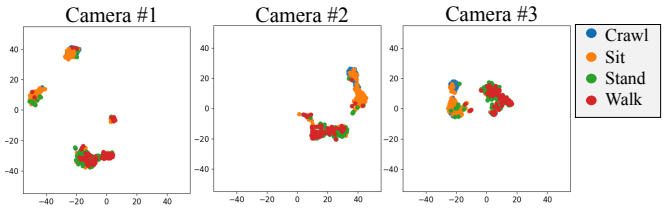
F. Action classification results

Results of our multiview action classification system integrating information from 5 different cameras are presented in the last column of Table IV, along with ablation results for the best three cameras with the baseline action classification methods described in previous section. Based on the ablation results, the proposed fusion strategy not only outperforms the best single-view classifier (81.3% vs 75.6% in the Leave One Super-session Out cross-validation and 79.1% vs 68.2% in the Leave One Subject Out cross-validation), but also outperforms all the other fusion strategies. As expected, combining information from different views in all cases results in similar or better results than using the best single camera. However, there is a significant improvement between the best post-processing strategy (sum logits) and the proposed approach (from 78.8% to 81.3% in the Leave One Super-session Out case, and from 72.3% to 79.1% in the Leave One Subject Out case). The difference observed between joint training of "sum logits" and "fused logits" (80.2% vs 81.3%, and 76.9% vs 79.1%) suggests that the improvement is the combined effect of joint training and the introduction of learnable fusion coefficients.

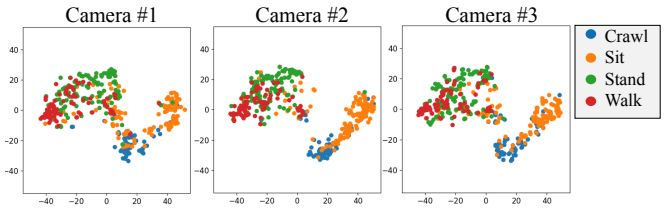
Low-dimensional representations of the spatio-temporal features extracted by the LSTM encoder $f_n \in \mathbb{R}^H$ were plotted for the three best views when: (a) view-specific LSTM encoders are trained, and (b) the proposed view-independent LSTM encoder is trained. In both cases the t-Distributed

Splits	Single view			Post-processing fusion			Joint training fusion		
				Three views			Three views		Five views
	#1	#2	#3	max score	mean score	sum logits	sum logits	fused logits	fused logits
Super-session 1	69.0 (1.9)	75.3 (1.4)	71.8 (0.6)	78.0 (1.1)	78.3 (1.1)	78.9 (1.1)	81.9 (0.2)	82.4 (0.4)	82.5 (0.2)
Super-session 2	70.7 (0.8)	75.4 (1.3)	73.4 (1.2)	77.0 (1.3)	76.7 (1.1)	78.7 (1.2)	79.1 (1.6)	82.2 (0.5)	83.6 (0.3)
Super-session 3	72.8 (2.4)	75.6 (0.3)	71.8 (0.5)	76.5 (1.2)	74.7 (1.1)	77.3 (0.8)	76.9 (0.2)	77.4 (0.2)	81.5 (0.5)
Super-session 4	66.4 (0.7)	76.1 (0.2)	74.9 (1.0)	79.3 (1.7)	78.7 (1.5)	80.2 (1.5)	83.1 (0.8)	83.1 (0.1)	84.2 (0.8)
Average	69.7 (0.8)	**75.6 (0.5)**	73.0 (0.4)	77.7 (0.7)	77.1 (0.6)	**78.8 (0.6)**	80.2 (0.5)	**81.3 (0.2)**	**83.0 (0.3)**
Subject T2	55.6 (1.1)	63.7 (0.9)	68.1 (0.6)	64.3 (1.6)	66.9 (1.2)	65.8 (1.5)	76.6 (0.4)	76.9 (1.2)	75.9 (0.6)
Subject T3	63.9 (2.6)	69.0 (2.0)	71.6 (1.1)	76.0 (1.9)	78.0 (1.6)	78.6 (1.7)	81.9 (2.0)	82.4 (1.2)	84.6 (0.9)
Subject T4	65.2 (2.2)	68.7 (2.1)	63.2 (1.9)	70.1 (2.4)	69.7 (2.0)	71.6 (2.1)	75.9 (1.0)	78.7 (1.1)	76.8 (0.5)
Subject T5 & T1	63.8 (1.9)	71.1 (1.5)	59.9 (2.0)	70.7 (1.7)	71.0 (1.2)	72.4 (1.9)	68.9 (2.4)	73.2 (0.4)	77.0 (0.7)
Subject T6	73.9 (2.4)	68.7 (4.7)	66.5 (2.4)	72.1 (1.7)	68.1 (2.7)	73.2 (2.9)	81.2 (1.0)	84.4 (1.0)	85.6 (0.5)
Average	64.5 (0.9)	**68.2 (1.2)**	65.8 (0.8)	70.7 (1.0)	70.7 (0.9)	**72.3 (0.8)**	76.9 (0.7)	**79.1 (0.5)**	**80.0 (0.3)**

TABLE IV: Action classification accuracy (%). Mean classification accuracy (standard deviation) from three runs is reported. Last column shows the results of the proposed method trained on 5 cameras. Ablation experiments performed in data from 3 best cameras are reported.



(a) t-SNE of clip embeddings from view-specific classifiers.



(b) t-SNE of clip embeddings from the proposed view-independent system.

Fig. 5: Example of t-SNE visualization [37] of 512-dimension LSTM clip embeddings of: (a) action classifiers trained specifically for each view, and (b) the proposed view-independent action classifier.

Stochastic Neighbor Embedding (t-SNE) [37] for visualization purposes was learned in an unsupervised way using data from the three views. An example of such plot for test data of Super-session 1 is shown in Fig. 5. As expected, spatio-temporal features are located in completely different regions when each encoder is trained independently. On the other hand, spatio-temporal feature embeddings generated by our method from different views are mapped to the same region, supporting the hypothesis that our method generates close to view-invariant features. Moreover, when data only from a single view are used at test time, the performance of the best single view classifier is significantly boosted (around 2% for Leave One Super-session Out and 6% for the Leave One Subject Out; see Table V). Thus, learning a (close to) view-invariant mapping improves performance even with missing views at test time.

Looking closer at the second-to-last column in Table IV, we note that the lowest action classification accuracy is achieved in splits where data from subject T1 are tested ("Super-session 3" and "Subject T5 & T1"). If we further analyze these splits it turns out that average classification accuracy of data from subject T1 is 43.0% in the "Super-session 3" split, and 47.9% in the "Subject T5 & T1" split (accuracy for the remaining data from these splits is close to the performance of the other splits: 82.9% and 80.6%, respectively). This could be explained by differences in data resolution (608× 808 for cameras #2 and #3 of T1 data, and 1080×1920 for other cameras and subjects), subject's age (T1 is the oldest subject), and subject's diagnosis (T1 is the only subject with Down Syndrome in the study). When all 5 cameras are used (last column of Table IV), the classification accuracy obtained in subject T1 data increases to 68.5% and 73.9% , respectively. This hints that the resolution discrepancy in T1 data for cameras #2 and #3 is responsible for part of the performance decrease. However, more data are needed to further investigate these hypotheses.

Splits	Single view		
	#1	#2	#3
Super-session 1	70.8 (0.6)	77.7 (0.9)	73.2 (0.8)
Super-session 2	70.8 (0.9)	78.1 (0.6)	73.4 (1.3)
Super-session 3	73.2 (1.3)	77.9 (0.9)	73.9 (1.6)
Super-session 4	69.1 (1.7)	76.0 (0.8)	74.6 (0.9)
Average	71.0 (0.6)	**77.5 (0.4)**	73.8 (0.6)
Subject T2	62.7 (0.8)	70.7 (1.1)	71.3 (0.5)
Subject T3	62.8 (3.9)	76.4 (0.9)	72.5 (3.8)
Subject T4	70.0 (1.6)	72.3 (1.7)	69.7 (1.4)
Subject T5 & T1	63.2 (1.9)	77.2 (1.8)	66.1 (2.3)
Subject T6	76.8 (0.7)	75.7 (2.8)	73.8 (1.5)
Average	67.1 (1.0)	**74.5 (0.8)**	70.7 (1.0)

TABLE V: Action classification accuracy (%). Considers LSTM encoder jointly trained for three views, but only data from one view used at test time; (cf. three first columns of Table IV).

Finally, we compare with earlier approaches to multi-view children activity recognition. Note that [7] reports that dense trajectories (DT) outperform CNNs for children activity recognition, and [8] proposes a DT-BOVW approach to which multiple instance learning SVM (MI-SVM) shows superior performance in [10]. As shown in Table VI, our method outperforms the MI-SVM approach by at least 10.8% in terms of average classification accuracy, even when MI-SVM uses information from all 5 cameras while our method uses only data from 3 of them. Obtaining such an improvement by the use of detection annotations for training suggests that spatial localization of the action is greatly advantageous in this

Splits	MI-SVM [10] Five views	Ours Three views	Ours Five views
Split 1	69.7 (1.0)	79.2 (0.9)	79.5 (0.3)
Split 2	72.8 (1.0)	78.8 (0.7)	81.2 (0.2)
Split 3	65.9 (0.4)	82.7 (0.9)	81.8 (0.5)
Average	69.4 (0.5)	**80.2 (0.5)**	**80.8 (0.2)**

TABLE VI: Action classification accuracy (%) on splits described in Table II, reported as Mean (std) from three runs.

challenging environment. Another factor possibly contributing to this performance increase is the use of deep features from ResNet-101 and LSTM as opposed to handcrafted DT features.

V. Conclusion

A view-independent detection-based multiview approach for action classification in infants suitable for complex scenes is presented. Such a method significantly outperforms the baseline model available for this dataset, even when fewer cameras are used. Departing from earlier approaches, we use pretrained deep features and detection annotations for training; both of which may have boosted performance. Experimental results show that the proposed method generates almost view-invariant features, which improves performance even when views are missing at test time. Finally, the introduction of learnable fusion coefficients based on detection scores was shown to be effective for combining information from different views, outperforming alternative fusion strategies.

Acknowledgment

This work has been supported by NIH #5R01HD87133.

References

[1] E. E. Aksoy, Y. Zhou, M. Wächter, and T. Asfour, "Enriched manipulation action semantics for robot execution of time constrained tasks," in *IEEE-RAS*, 2016, pp. 109–116.

[2] D. Singh and C. K. Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognit.*, vol. 65, pp. 265 – 272, 2017.

[3] Y. Gao *et al.*, "Human action monitoring for healthcare based on deep learning."

[4] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *CVPR*, 2019, pp. 7882–7891.

[5] J. Goto, T. Kidokoro, T. Ogura, and S. Suzuki, "Activity recognition system for watching over infant children," in *RO-MAN*.

[6] T. L. Westeyn, G. D. Abowd, T. E. Starner, J. M. Johnson, P. W. Presti, and K. A. Weaver, "Monitoring children's developmental progress using augmented toys and activity recognition," *Personal and Ubiquitous Computing*, vol. 16, no. 2, pp. 169–191, 2012.

[7] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi-view fusion for action recognition in child-robot interaction," in *ICIP*, 2018, pp. 455–459.

[8] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots," in *ICRA*, 2018, pp. 1–8.

[9] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," *arXiv:1812.00253 [cs.RO]*, 2018.

[10] E. Kokkoni, E. Mavroudi, A. Zehfroosh, J. C. Galloway, R. Vidal, J. Heinz, and H. G. Tanner, "Gearing smart environments for pediatric motor rehabilitation," *J Neuroeng Rehabil*, vol. 17, no. 1, p. 16, 2020.

[11] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, "On the estimation of children's poses," in *ICIAP*, 2017, pp. 410–421.

[12] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402 [cs.CV]*, 2012.

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv:1705.06950 [cs.CV]*, 2017.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV*, 2011.

[15] S. Suzuki, Y. Mitsukura, H. Igarashi, H. Kobayashi, and F. Harashima, "Activity recognition for children using self-organizing map," in *RO-MAN*, 2012, pp. 653–658.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeuIPS*, 2015, pp. 91–99.

[17] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[18] S. Boughorbel, J. Breebaart, F. Bruekers, I. Flinsenberg, and W. Ten Kate, "Child-activity recognition from multi-sensor data," in *Mesuring Behavior*, 2010, pp. 1–3.

[19] A. Hosseini, S. Fazeli, E. van Vliet, L. Valencia, R. Habre, M. Sarrafzadeh, and A. Bui, "Children activity recognition: Challenges and strategies," in *EMBC*, 2018, pp. 4331–4334.

[20] A. Hosseini, D. Zamanzadeh, L. Valencia, R. Habre, A. A. Bui, and M. Sarrafzadeh, "Domain adaptation in children activity recognition," in *EMBC*, 2019, pp. 1725–1728.

[21] D. Karch, K.-S. Kang, K. Wochner, H. Philippi, M. Hadders-Algra, J. Pietz, and H. Dickhaus, "Kinematic assessment of stereotypy in spontaneous movements in infants," *Gait & posture*, vol. 36, no. 2, pp. 307–311, 2012.

[22] C.-Y. Fang, H.-H. Hsieh, and S.-W. Chen, "A vision-based infant respiratory frequency detection system," in *DICTA*, 2015, pp. 1–8.

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

[24] T. Tsuji *et al.*, "Markerless measurement and evaluation of general movements in infants."

[25] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[26] X. Peng and C. Schmid, "Multi-region two-stream r-cnn for action detection," in *ECCV*, 2016, pp. 744–759.

[27] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *CVPR*, 2018, pp. 1130–1139.

[28] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *ICCV*, 2019, pp. 6212–6221.

[29] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *CVPR*, 2015, pp. 2458–2466.

[30] Y. Kong, Z. Ding, J. Li, and Y. Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 3028–3037, 2017.

[31] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[32] E. Kokkoni, S. W. Logan, T. Stoner, T. Peffley, and J. C. Galloway, "Use of an in-home body weight support system by a child with spina bifida," *Pediatr. Phys. Ther.*, vol. 30, no. 3, pp. E1–E6, 2018.

[33] E. Kokkoni and J. C. Galloway, "User-centred assistive technology assessment of a portable open-area body weight support system for in-home use," *Disabil. Rehabil.: Assist. Technol.*, pp. 1–8, 2019.

[34] L. A. Prosser, L. B. Ohlrich, L. A. Curatalo, K. E. Alter, and D. L. Damiano, "Feasibility and preliminary effectiveness of a novel mobility training intervention in infants and toddlers with cerebral palsy," *Dev. Neurorehabilit.*, vol. 15, no. 4, pp. 259–266, 2012.

[35] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *CVPR*, 2017, pp. 7310–7311.

[36] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580 [cs.NE]*, 2012.

[37] L. v. D. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.