Video-Based Activity Recognition for Infants

Keyuan Wu
Stanford University
Stanford, CA
keyuanwu@stanford.edu

Abstract

Infants under 12 months old can be vulnerable as they are unable to communicate verbally. Utilizing video cameras to recognize their actions and send alerts to caregivers is of utmost importance. While extensive research has been conducted on action recognition for adults, such as sports activities, the study on action recognition for babies remains limited. Infants primarily engage in a limited range of actions, including crawling, eating, crying, smiling, playing, and beginning to walk and dance, with occasional falls. Consequently, the problem becomes more challenging due to the reduced variety of actions. However, with the rapid advancements in computer vision and video understanding, numerous models have emerged, particularly with the successful application of transformers in computer vision tasks. In this paper, I have collected a dataset of infant action videos from open sources, encompassing 10 distinct classes of activities. I employed transfer learning techniques on state-of-the-art pretrained transformer-based models, such as Vision Transformer (ViT), Swin Transformer, Video Vision Transformer (ViViT), and Video Swin Transformer, for recognizing and predicting baby activities. Video durations ranged from 1 to 10 seconds. My experiments revealed that treating the video as a stack of images and applying ViT offers a robust baseline, achieving an accuracy of 43.33% for top-1 prediction. The best-performing model involved transfer learning from Swin Transformer, combining LSTM (Long Short-Term Memory) to perform late fusion, yielding an accuracy of 50.00% for top-1 prediction. Despite these advancements, the overall top-1 accuracy still remains relatively low. This can be attributed to multiple factors, including an inadequate sample size for training, a lack of data augmentation techniques, and suboptimal utilization of temporal information. To further improve the accuracy and effectiveness of the models, addressing these challenges is essential.

1. Introduction

Video understanding is a subfield of computer vision that focuses on extracting meaningful information and comprehending the content and context within videos.Unlike static images, videos provide rich temporal information, including motion, object interactions, and scene dynamics. By leveraging this temporal aspect, video understanding algorithms aim to recognize and interpret various visual phenomena such as actions, events, object tracking, object recognition, and semantic understanding. The field of video understanding has witnessed significant advancements in recent years, largely driven by deep learning techniques and the availability of large-scale annotated video datasets. State-of-the-art techniques, such as two-stream networks, 3D convolutional networks, LSTM networks, Transformer-based models, and self-supervised learning, have significantly advanced the capabilities of machines in recognizing actions, events, and other visual phenomena within videos. These advancements open up exciting possibilities for applications such as video surveillance, autonomous vehicles, video editing, and augmented reality.

Action recognition is a fundamental task in computer vision and video understanding that aims to automatically analyze and understand human actions or activities depicted in videos. It involves developing algorithms and models capable of recognizing and categorizing different types of actions, such as walking, running, jumping, and interacting with objects. Action recognition plays a crucial role in a wide range of applications, including video surveillance, human-computer interaction, sports analysis, autonomous systems, and video content analysis.

This paper studies action recognition for infants with age less than 12 months. The idea is inspired by the papers of Hammal et al. [10] and Olalere [9]. Hammal pointed out that most of the available open source or commercial action unit (AU) detectors are trained with the faces of young adults. Infant faces differ from adult faces in terms of proportion (e.g. larger eyes and smaller jaw-to-face ratio), skin smoothness, amount of texture and wrinkles and presence of brow knitting action. For these reasons, AU detectors

trained with adult faces may not be well suited to detect actions in infant faces. Models specifically trained to detect AUs in infant faces are needed. Olaere in his paper tested kid specific model and adult specific model and concluded that the kid-specific dataset is more complex to generalize to than the adult-specific dataset. The study also shows that the features learned from training on a kid-specific dataset alone can be used to classify adult activities while the reverse is not the case.

Moreover, infants under 12 months of age are unable to speak and may be at risk if they are not within the caregiver's line of sight. Therefore, the objective of this study is to develop a network that can accurately recognize different baby actions, regardless of whether they occur under adult supervision or not. To accomplish this, a dedicated dataset for infant actions has been created, encompassing various activities such as crawling, crying, falling down and walking. To achieve state-of-the-art performance, a selection of transformer-based models has been fine-tuned for this specific task.

2. Related Work

Hammal et al. [10] primarily focuses on detecting facial expressions in infants, specifically actions such as raising the inner/outer corner of the eyebrow, raising the cheeks, pulling up the lip corners orthogonally, and pulling the lip corners laterally. The study employs an 8-layer multi-label CNN network comprising 5 convolutional layers and 3 fully connected layers.

Olalere [9] employed the classes from the Kinetics-400 dataset. The kinetic-kids dataset contains clips of kids within the ages 0-12 years old performing 21 sporting activities. The kinetic-adults dataset comprises the same activity classes but with adult participants. The author conducted a comparison between the kid model, adult model, and mixed model. The findings indicated that the kid-specific dataset is more complex to generalize to than the adult-specific dataset.

Khan [6] focuses on studying baby sleeping postures, with particular emphasis on addressing potentially dangerous situations where the baby's face is covered or when the baby throws off the blanket during sleep. The authors also place significant emphasis on the development of the device and hardware, as well as the design of the system to effectively send notifications and generate alerts to the caregiver's smartphone. The primary model utilized in this study is the Multi-Task Convolutional Neural Network (CNN).

Sujitha Balasathiya et al. [3] investigates a diverse range of activities including crawling, crying, laughing, sleeping, standing, playing, sitting, eating, and walking. The authors explore various handcrafted feature extraction techniques, such as Histogram of Oriented Gradients (HOG) features. The classification algorithms employed include Multi-class Naive Bayes, Support Vector Machine, Ensemble classifier, Discriminant analysis, and Decision tree classifiers. The study does not utilize neural network models for the action recognition task.

Dechemi et al. [4] investigates infant reaching action recognition. The authors conducted a comparative analysis of various network structures, including ResNet (without data augmentation), ResNet (with data augmentation), ResNet + LSTM, and O-LSTM (a single layer LSTM trained by the optical flow images).

Adewopo et al. [1] aligns closest with my research interest due to the modern techniques they used. The authors employ transfer learning using the I3D model as a benchmark and compare it with I3D-CovLSTM2D models with and without Video Augmentation. They collected their own dataset from open-source videos sourced from social media platforms. The dataset focuses on activities such as baby climbing the crib, moving out of the crib, and getting stuck.

The previous studies have certain limitations that prevent them from fully serving my research objectives. Sujitha Balasathiya et al. [3] covers a diverse range of activities, but it lacks modern neural network techniques. Olalere [9] also includes diverse activities but focuses on children aged 0-12 years old, specifically studying their involvement in sports activities. Hammal et al. [10], Khan [6], Dechemi et al. [4], and Adewopo et al. [1] have specific and narrow conditions, such as analyzing sleeping gestures [6], crib-related sleeping and awakening [1], facial analysis [10], or reaching actions [4]. In contrast, my study encompasses a diverse set of activities similar to [3], but specifically for infants under one year old. In terms of model development, I have applied a number of transformer-based models. I have collected a dataset dedicated to baby-specific activities from YouTube. It is important to note that my paper will not cover hardware or alert system development. While critical moments of detecting a baby in danger may require specific models, similar to papers [6] or [1], my paper serves the general purpose of recognizing various baby actions.

3. Methods

3.1. Model I. Transfer Learning Vision Transformer (ViT)

Dosovitskiy et al. [5] introduced the Vision Transformer (ViT) architecture, which employs self-attention mechanisms to process image data. Unlike CNNs that rely on convolutions, ViT divides the input image into fixed-size patches and maps them to sequences of tokens. This allows the application of Transformer principles, enabling the model to capture long-range dependencies and leverage global context information.

To suit my specific requirement of 10 target classes, I uti-

lized the pre-trained ViT model and adjusted the number of outputs in the fully connected layers accordingly. Although the ViT model is primarily designed for image classification, I modified its usage to handle videos by treating them as sequences of individual frames. To accomplish this, I extracted every frame from the video and created an image dataset by selecting every 10th frame, associating each frame with the corresponding label obtained directly from the video.

During the data preprocessing stage, I applied standard transformations to the images, resizing them to 224x224 pixels. Additionally, I normalized the images using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]). I used a batch size of 8.

For predicting the video labels, I experimented with two methods. In the first method, I selected the label that appeared most frequently among the frames, using a majority vote approach. In the second method, I identified the frame with the highest predicted probability and used the label from that particular frame as the video label.

3.2. Model II - Transfer Learning Swin transformer + LSTM (late fusion)

Liu et al. [7] introduced the Swin Transformer, which utilizes a shifted window-based self-attention mechanism. The Swin Transformer performs self-attention computations locally within non-overlapping windows. To incorporate global contexts, these windows are shifted across different blocks in the transformer, enabling interactions across the entire image space. One notable feature of the Swin Transformer is its ability to operate directly on image patches instead of token sequences. This approach facilitates efficient and flexible modeling of visual data, leading to impressive performance across various vision tasks.

In the context of video processing, a common strategy involves analyzing individual frames or video chunks separately and then combining these analyses at the end of the model to make a final prediction. For instance, one could apply a Swin Transformer to each frame or small sequence of frames in the early part of the model. Subsequently, a LSTM or another suitable method can be employed to fuse the frame-wise predictions or features in the later part of the model. This approach captures temporal dependencies between frames and generates a comprehensive prediction for the entire video. This ”late fusion” approach helps the model capture diverse information present in the data and often leads to a more robust and accurate model.

During the data preprocessing phase, I extracted frames from each video. To ensure a representative sampling of frames and to avoid bias towards specific sections of the video, a random selection process was employed. Specifically, 16 frames were randomly chosen from the available frames within each video. To enhance model generalization, the inclusion of a dropout layer will be determined later during the experimentation phase. Additionally, the output features of the linear layer were adjusted to match the desired 10 classes. Standard normalization techniques were employed, including resizing the frames to 224x224 pixels and normalizing them using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]. A batch size of 8 was used.

3.3. Model III - ConvNet (early fusion)

In contrast to late fusion approach, early fusion involves combining information from multiple sources or time points early on in the model architecture, allowing for the integration of spatial and temporal dependencies from the beginning.

The early fusion architecture begins by extracting frames from each video and applying frame-level fusion. 16 frames are selected, through a random sampling process, to ensure representation from different time points throughout the video. These frames are then fused together in the early stages of the model, combining their spatial information.

The fused representation, which captures both spatial and temporal aspects of the video data, is passed through a convolutional neural network (ConvNet). The ConvNet consists of multiple layers, including convolutional layers, ReLU activation functions, and max pooling layers. These layers are designed to extract high-level visual features and downsample the spatial dimensions of the input. The output of the ConvNet is then flattened and fed into a fully connected layer, which maps the features to 10 classes.

During the data preprocessing phase, standard normalization techniques were applied to the frames. This involved resizing the frames to a consistent size of 224x224 pixels, ensuring uniformity in the input data. Additionally, the frames were normalized using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]. A batch size of 8 was used.

3.4. Model IV - Transfer Learning Video Vision Transformer (ViViT)

Video Vision Transformer (ViViT) [2] is a model architecture designed to extend the capabilities of the Vision Transformer (ViT) for video analysis tasks.

To classify video samples using ViViT, the authors propose two methods for embedding video samples into the model. The first method, Uniform Frames Sampling, is similar to ViT's patch-based approach for images. Each frame in the video is divided into patches, which are treated as tokens. However, this approach doesn't capture the exact frame and time index of each patch within the video sample. The second method, Tubelet Embedding shown in Figure 1, addresses the limitations of Uniform Frames Sampling. Instead of extracting patches from each frame, the authors suggest a new type of token called a ”tubelet.” A series of patches is extracted from a video clip, forming a

tubelet. Each token in this approach captures both spatial and temporal information, providing a more comprehensive representation of the video sample.

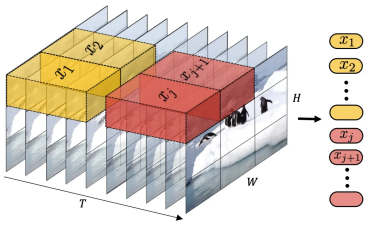

Figure 1. ViViT: Tubelet Embedding

In the paper, the author proposes four different variants of pure transformer-based models for video classification, inspired by ViT. Among all these models, Model 2: Factorized Encoder-Decoder demonstrates the best performance. Illustrated in Figure 2, this model introduces the concept of factorization. The video is divided into small clips, with each clip processed by a spatial transformer. The resulting encoding vectors, along with a CLS token, are combined with position embeddings and passed through a temporal transformer encoder. The temporal CLS token is then subjected to an MLP head for classification. The spatial transformer operates by treating each token as a Tubelet extracted from one clip, where all tokens originate from the same temporal index but different spatial indexes. On the other hand, the temporal transformer processes each token as a vector extracted from a clip, implying that each token corresponds to a distinct temporal index within the video.

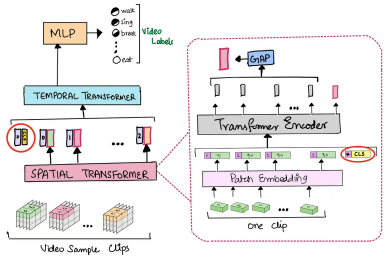

Figure 2. ViViT: Model-2: Factorized Encoder-Decoder

For my task, I employed Model 2 as the transfer learning model. I modified the linear layer's number of output features to 10. In terms of data preprocessing, I standardized the frame sizes by resizing them to 224x224 pixels. To ensure uniform normalization, I divided the frame tensor by 255. To maintain a balanced representation during both training and evaluation, I uniformly selected 16 frames from each video.

3.5. Model V - Transfer Learning Video Swin Transformer

Video Swin Transformer [8] introduces a locality inductive bias to the self-attention module, allowing for more efficient and effective video recognition.

The architecture of Video Swin Transformer shown in Figure 3 includes a 3D shifted window-based MSA module and a feed-forward network (FFN) within each block. Layer Normalization is applied before each MSA module and FFN, and residual connections are utilized. The output features are computed using regular or shifted window partitioning configurations. Furthermore, the architecture incorporates a 3D relative position bias for each head in the self-attention computation, which aids in capturing positional information and enhancing the model's performance.

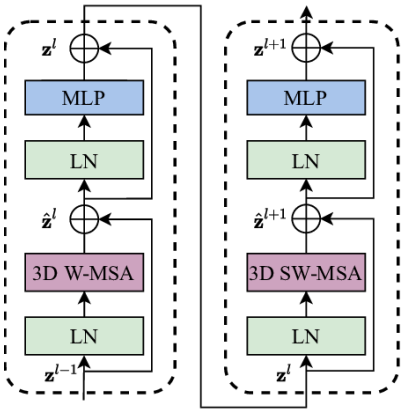

Figure 3. Video Swin Transformer blocks

W-MSA (Window-based Multi-head Self-Attention) introduces the concept of windows as an extension to the Multi-head Self-Attention (MSA) mechanism. The video is partitioned into non-overlapping 3D windows, arranged in a partitioned manner, and MSA is applied within each window to capture spatial and temporal dependencies.

To further enhance representation power and introduce cross-window connections, the author proposes SW-MSA (Shifted Window-based Multi-head Self-Attention). This approach extends the shifted window mechanism from the Swin Transformer, originally designed for 2D images, to 3D windows. Figure 4 provides an illustration of the 3D shifted windows.

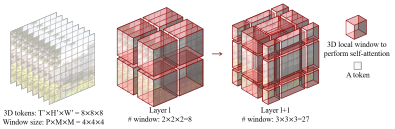

Figure 4. An illustrated example of 3D shifted windows

Moreover, this model is adapted from Swin Transformer for image recognition, allowing it to leverage the strength of strong pre-trained image models.

In my experiments, I selected Swin-T as the model for transfer learning. Similar to Transfer Learning ViViT, I adjusted the number of output features in the linear layer to 10. Additionally, I resized the frames to 224x224 pixels and normalized the frame tensor by dividing it by 255.

4. Data

Initially, I gathered a dataset of 82 raw videos sourced from YouTube, ranging in duration from 10 seconds to 10 minutes. Some of the longer videos consisted of compilations of shorter videos, which I trimmed and edited using HitPaw Video Converter. In certain cases, cropping was applied to focus on specific content.

After the editing process, the video durations varied from 1 second to 2 minutes. However, during the training phase of the model, I observed that the 2-minute videos were too lengthy for data preprocessing. Consequently, I made the decision to retain videos with a duration of less than 15 seconds, resulting in an average duration of approximately 5 seconds.

Each video in the dataset was labeled with specific categories, including crawling, crying, dancing, eating, falling slide, falling, playing with animals, playing with toys, smiling, and walking. In total, there were 10 distinct categories. The number of videos and average length for each category are presented in Table 1. I captured a screenshot from a randomly selected video in each category and displayed them in Figure 3.

Following the preprocessing and labeling stages, I obtained the finalized dataset, which would be utilized for subsequent analysis and modeling tasks.

Table 1. Summary of Video Statistics

Category	Number of Videos	Average Length
crawling	29	6.36s
crying	35	6.87s
dancing	21	8.16s
eating	21	11.59s
falling slide	39	4.77s
falling	41	3.74s
playing with animals	25	7.77s
playing with toys	25	6.92s
smiling	26	7.93s
walking	43	4.17s

5. Results

I have a total of 305 videos, which I divided into training, testing, and validation sets using a ratio of 7:2:1. The training set consisted of 213 videos, the testing set consisted of 62 videos, and the validation set consisted of 30 videos. To ensure consistency, I set a random seed, which guarantees that each model will have the same training, testing, and validation datasets.

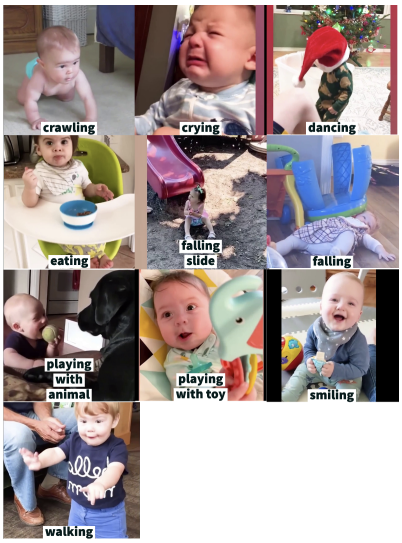

Figure 5. Screenshorts of Sample Videos

During training, both the training and testing samples were used to monitor loss and accuracy. After finalizing the model, I evaluated its performance using the independent validation set. The evaluation involved comparing top-1 and top-5 accuracies across different models.

For all models, I experimented with learning rates of 1e-3, 1e-4 or 1e-5, and I maintained a batch size of 8. The number of training epochs varied in the range of 10 to 100. To calculate the loss during training, I used the CrossEntropy loss function. For optimization, I employed the Adam optimizer. The specific configurations for learning rate decay and dropout ratio varied depending on the model.

In terms of data preprocessing, in addition to resizing and normalization, for each video, I always selected 16 frames evenly distributed throughout the duration.

Due to time constraints, I was unable to incorporate data augmentation techniques into the training process, although data augmentation is known to enhance model accuracy by introducing variations in the training data.

5.1. Model I. Transfer Learning Vision Transformer (ViT)

This model is a pre-trained Vision Transformer (ViT), specifically the "vit_base_patch16_224" variant. The fully connected layer of the model was modified to accommodate the output of 10 classes. The weights of the pre-trained model were frozen, ensuring that they remain unchanged during the training process. To address this issue and en-

hance generalization, a dropout layer was introduced before the fully connected layers. Through experimentation, I tested dropout rates of 0.2 and 0.5, ultimately settling on a rate of 0.5 for the final model configuration. Figure 6 illustrates the loss and accuracy metrics for both training and testing datasets.

In terms of performance evaluation, two different approaches were employed. The first approach involved selecting the most frequent label among all frames. This method achieved a top-1 accuracy of 43.33% and a top-5 accuracy of 93.33%. Alternatively, the second approach focused on selecting the label with the highest probability. With this method, the top-1 accuracy improved to 46.67%, and the top-5 accuracy increased to 56.67%. Although the first approach demonstrated slightly lower top-1 accuracy, it outperformed the second approach in terms of top-5 accuracy. As a result, I considered the performance of method 1 as my baseline model. However, it is noteworthy that category of dancing, playing with toys, and smiling exhibited a prediction accuracy of 0.00% in validation dataset for this model.

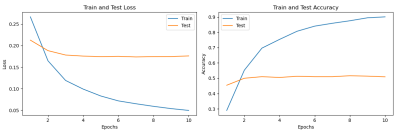

Figure 6. Model I: Loss and Accuracy

5.2. Model II - Transfer Learning Swin transformer + LSTM (late fusion)

This model comprises three key components: a feature extractor, an LSTM (Long Short-Term Memory) layer, and a fully connected layer. The feature extractor is based on the pre-trained Swin Transformer model, specifically the "swin_tiny_patch4_window7_224" variant. The weights of the pre-trained model were frozen, ensuring that they remain unchanged during the training process. The LSTM layer takes the output of the feature extractor, which has a size of 768, and generates a sequence of hidden states with a size of 256. The fully connected layer operates on the last output of the LSTM layer. Firstly, it applies a linear transformation to reduce the dimensionality of the input from 768 to 64. Then, a ReLU activation function is applied to introduce non-linearity. To prevent overfitting, a dropout layer with a dropout probability of 0.4 is included. Finally, another linear transformation maps the 64-dimensional output to the 10 classes of the classification task.

During experimentation, two different learning rates, 1e-4 and 1e-5, were tested. Based on the results, a learning rate of 1e-4 was selected for the final model. The training process was initially extended to 50 epochs, but after careful evaluation, it was determined that a more optimal choice was 10 epochs. Weight decay (L2 regularization) was also explored, but ultimately not utilized in the final model configuration. Figure 7 illustrates the loss and accuracy metrics for both training and testing datasets.

With these adjustments, the final model achieved a top-1 accuracy of 50.00% and a top-5 accuracy of 83.33%. However, certain categories, such as dancing, eating, and playing with toys, still exhibited a prediction accuracy of 0.00% in the validation dataset.

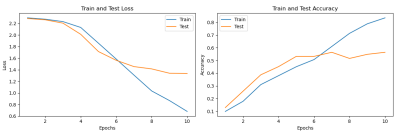

Figure 7. Model II: Loss and Accuracy

5.3. Model III - ConvNet (early fusion)

This model is a Convolutional Neural Network (ConvNet) specifically designed for image classification tasks. It comprises several convolutional layers followed by ReLU activation and max pooling layers.

During the training phase, it was observed that training for 20 epochs resulted in overfitting. Consequently, I opted to train the model for 10 epochs. Additionally, I decided to incorporate weight decay with a value of 0.001 in the final model configuration. A learning rate of 1e-4 was employed for training. Figure 8 illustrates the loss and accuracy metrics for both training and testing datasets.

With these modifications, the final model achieved a top-1 accuracy of 43.33% and a top-5 accuracy of 86.67%. Dancing and playing with toys still exhibited a prediction accuracy of 0.00% in the validation dataset.

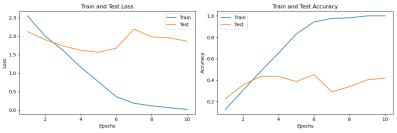

Figure 8. Model III: Loss and Accuracy

5.4. Model IV - Transfer Learning Video Vision Transformer (ViViT)

The model is a pre-trained Video Vision Transformer (ViViT). The fully connected layer was modified to output probabilities for 10 classes. The weights of the pre-trained model were frozen to maintain their original values during training.

I tested various learning rates including 1e-3, 1e-4, and 1e-5, in an attempt to find the optimal value. However, the model consistently exhibited underfitting, with both the

training and testing accuracies remaining low. After training for 50 epochs, the training accuracy reached 36.15%, while the testing accuracy was only 25.81%. No dropout or weight decay techniques were applied. Figure 9 illustrates the loss and accuracy metrics for both training and testing datasets.

The final model achieved a top-1 accuracy of 33.33% and a top-5 accuracy of 80.00% on the validation dataset. Certain categories such as dancing, playing with toys, and smiling still presented a prediction accuracy of 0.00%.

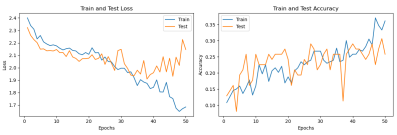

Figure 9. Model IV: Loss and Accuracy

5.5. Model V - Transfer Learning Video Swin Transformer)

The model is a pre-trained Video Swin Transformer, which was fine-tuned for a 10-class classification task. The weights of the pre-trained model were frozen, ensuring that they remained fixed during the training process.

Various learning rate configurations were tested during the training process, including 1e-3, 1e-4 and 1e-5, with and without weight decay. After 100 epochs, the model achieved a training accuracy of 100.00%, but the testing accuracy was only 32.26%.

When evaluating the final model on a separate validation dataset, it achieved a top-1 accuracy of 36.67% and a top-5 accuracy of 66.67%. Certain categories such as dancing, eating, playing with animals, and smiling still exhibited a prediction accuracy of 0.00%. Figure 10 illustrates the loss and accuracy metrics for both training and testing datasets.

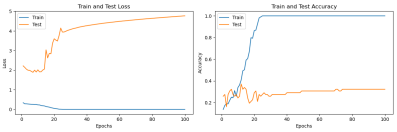

Figure 10. Model V: Loss and Accuracy

Table 2 displays the hyperparameters employed in the final model along with the corresponding model accuracies. Out of the five proposed models, Model II - Transfer Learning Swin Transformer + LSTM (late fusion) achieves the best results.

Table 3 presents the top-1 prediction accuracy for each model across 10 categories. It is important to note that the validation dataset was limited in size. Upon examining the videos associated with the ”dancing” and ”playing with toys” categories, it was observed that they can be easily confused with other categories, leading to a prediction accuracy of 0.00% across all models. Conversely, categories such as ”crawling,” ”crying,” and ”falling slide” exhibited relatively higher prediction accuracies, indicating that they are more distinguishable and predictable within the dataset.

6. Conclusion

In conclusion, this study conducted experiments on five transfer learning models based on Vision Transformer (ViT) and Swin Transformer pre-trained models. The best-performing model was the transfer learning Swin Transformer with late fusion LSTM, achieving an accuracy of 50.00% for top-1 prediction.

Among the ten categories, some were relatively easier to classify, such as crawling, crying, and falling down. However, certain categories proved to be extremely challenging, such as dancing and playing with toys. Extracting discrete images from the videos made it difficult to capture the nuances of dancing movements, as infants often appeared stationary. Additionally, the presence of various toys in the frames, after cropping, resulted in incomplete toy visibility, posing challenges for accurate prediction. Similar difficulties were encountered with the category of animals.

Several limitations should be acknowledged in this study. A total of 305 videos with an average length of 6.33 seconds and a cumulative duration of 32 minutes were collected. Due to time and resource constraints, I did not perform data augmentation, which could have provided additional data points and improved the model's accuracy. Additionally, I only trained the last linear layers of the models while keeping the remaining weights frozen. Allowing more time to fine-tune all the weights might have led to increased accuracy. Lastly, the performance of the video-based models ViViT and Video Swin Transformer fell short of expectations. Further investigation is required to identify potential reasons, such as exploring hyperparameter settings or optimizing the input data to maximize the utilization of the model architecture. Overall, this study supports the notion that ViT or Swin Transformer serves as a solid baseline for video classification. However, the video-based models exhibited subpar performance.

It is crucial to acknowledge that the average video length in this study was only 6.33 seconds, and the analysis focused solely on visual data. However, in reality, video cameras are often operational 24/7, necessitating the adoption of additional techniques to capture critical moments effectively. One such technique involves leveraging audio data, which can significantly enhance video surveillance capabilities. By incorporating audio information, it becomes possible to identify crucial or risky moments when infants may be in danger, complementing the visual data and providing a more comprehensive understanding of the context. This integration of audio and visual data can offer valuable in-

Table 2. Model Hyperparameters and Accuracies

Model	Learning Rate	Batch Size	Number of Epochs	Optimizer	Learning Rate Decay	Top1 Accuracy	Top5 Accuracy
Model I	1e-4	8	10	Adam	0.1	43.33%	93.33%
Model II	1e-4	8	10	Adam	-	50.00%	83.33%
Model III	1e-4	8	10	Adam	0.001	43.33%	86.67%
Model IV	1e-4	8	50	Adam	-	33.33%	80.00%
Model V	1e-4	8	100	Adam	-	36.67%	66.67%

Table 3. Top-1 Prediction Accuracies by Category

Label	Videos	Model I	Model II	Model III	Model IV	Model V
crawling	3	0.67	0.33	0.67	0.67	0.33
crying	3	0.67	0.67	0.67	0.00	0.33
dancing	2	0.00	0.00	0.00	0.50	0.00
eating	2	0.50	0.00	0.50	0.50	0.00
falling slide	4	0.75	0.75	0.50	0.25	0.50
falling	4	0.50	0.75	0.75	0.75	0.75
playing w animals	3	0.67	0.67	0.33	0.33	0.00
playing w toy	2	0.00	0.00	0.00	0.00	0.00
smiling	3	0.00	0.33	0.33	0.00	0.33
walking	4	0.25	0.75	0.25	0.25	0.25

sights for ensuring the safety and well-being of infants in real-world scenarios.

7. Contributions & Acknowledgements

I would like to acknowledge the following resources and materials that greatly assisted me in completing this project. While this project was entirely conducted by myself, the contributions of the following documents and repositories were instrumental in its successful completion:

The GitHub repository "ViViT-pytorch" (https://github.com/rishikksh20/ViViT-pytorch) provided valuable implementation details and code examples related to the ViViT (Video Vision Transformer) model.

The GitHub repository "Video-Swin-Transformer" (https://github.com/SwinTransformer/Video-Swin-Transformer) proved to be a valuable resource for information on the Swin Transformer model applied to video understanding.

The Medium article "ViViT: Video Vision Transformer" (https://medium.com/aiguys/vivit-video-vision-transformer-648a5fff68a4) provided an insightful and comprehensive overview of the ViViT model.

The YouTube channel of @Mu Li offered a wealth of educational content related to computer vision and transformer-based models.

References

[1] Victor Adewopo, Nelly Elsayed, and Kelly Anderson. Baby physical safety monitoring in smart home using action recognition system. 2022. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 3

[3] S. Sujitha Balasathiya, S. Mohamed Mansoor Roomi, and B. Sathyabama. Infant action database: A benchmark for infant action recognition in uncontrolled condition. *Journal of Physics: Conference Series*, 1917:012019, 2021. 2

[4] A. Dechemi et al. Babynet: A lightweight network for infant reaching action recognition in unconstrained environments to support future pediatric rehabilitation applications. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 461–467, Vancouver, BC, Canada, 2021. IEEE. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[6] Tareq Khan. An intelligent baby monitor with automatic sleeping posture detection and notification. *AI*, 2:290–306, 2021. 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 3

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 4

[9] F.E. Olalere. Video-based activity recognition for child behaviour understanding, 2021. 1, 2

[10] I. Onal Ertugrul, Y.A. Ahn, M. Bilalpur, et al. Infant afar: Automated facial action recognition in infants. *Behavior Research*, 55(4):1024–1035, 2023. 1, 2