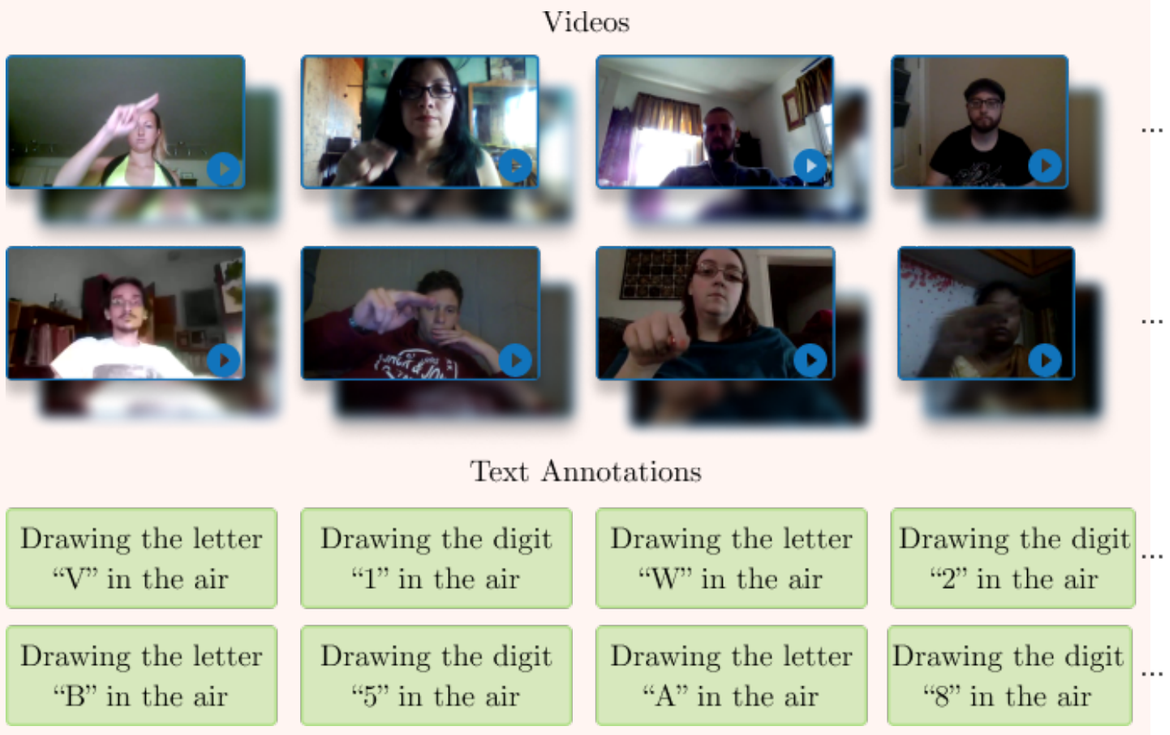
AirLetters: An Open Video Dataset of Characters Drawn in the Air

Rishit Dagli[1, 2] Guillaume Berger[1] Joanna Materzynska[3] Ingo Bax[1]
Roland Memisevic[1]

[1] Qualcomm AI Research [2] University of Toronto [3] MIT

Summary



- We introduce a new real-world dataset, utilizing human generated articulated motions with videos of people drawing Latin characters and labels.
- Unlike existing video datasets, accurate video understanding on our dataset requires detailed understanding of motion in the video and the integration of long-range information across the entire video.
- We show that existing image and video understanding models perform poorly and fall far behind the human baseline.

Dataset

Figure 1. **Diversity in our Dataset.** Each of the images is taken from a randomly sampled video from our dataset. Our dataset has a large variance in the appearance of subjects, background, occlusion, and lighting conditions in the videos.

- We focus on manual articulations of each letter of the Latin alphabet as well as numeric digits. This amounts to 36 primary gesture classes, for which recognition requires temporal and spatial analysis of the video.
- We also include two contrast classes designed to refine the sensitivity and specificity of recognition systems trained on our dataset. The "Doing Nothing" class includes videos of individuals in non-active states to represent periods of inactivity within human-computer interactions, and the "Doing Other Things" class consists of clips capturing miscellaneous non-communicative movements.

Table 1. **Dataset Statistics,** showing the number of classes, number of actors and median values for duration, frames per second (FPS), videos per class, and videos per actor.

Statistic	Value (Total)
Videos	161652
Classes	38
Actors	1781
Frames	40142100

Statistic	Value (Median, σ)
Duration	2.93 (±0.13)
FPS	30.0 (±0.0)
Videos per Class ($\times 10^3$)	4.04 (±1.31)
Videos per Actor	40.0 (±99.29)

Experiments

Table 2. **Classification accuracy** of multiple image models, video models, and (large) vision language models on the AirLetters dataset.

Method	Top-1 Acc (↑)
Image Models	
ViT-B/16	7.49
MaxViT-T	7.56
ResNet-50	13.87
Vision Language Models	
Video-LLaVA (w/o contrast)	2.53
VideoLLaMA2 (w/o contrast)	2.47
Video-LLaVA	7.29
VideoLLaMA2	7.58

Method	Top-1 Acc (↑)
Video Models	
VideoMAE (16)	57.96
ResNet-101 + LSTM	58.45
ResNet-50 + LSTM	63.24
Sense [69]	65.97
ResNext-152 3D	65.77
ResNext-101 3D	69.74
ResNext-200 3D	71.20
Human Performance (10 videos/class)	96.67

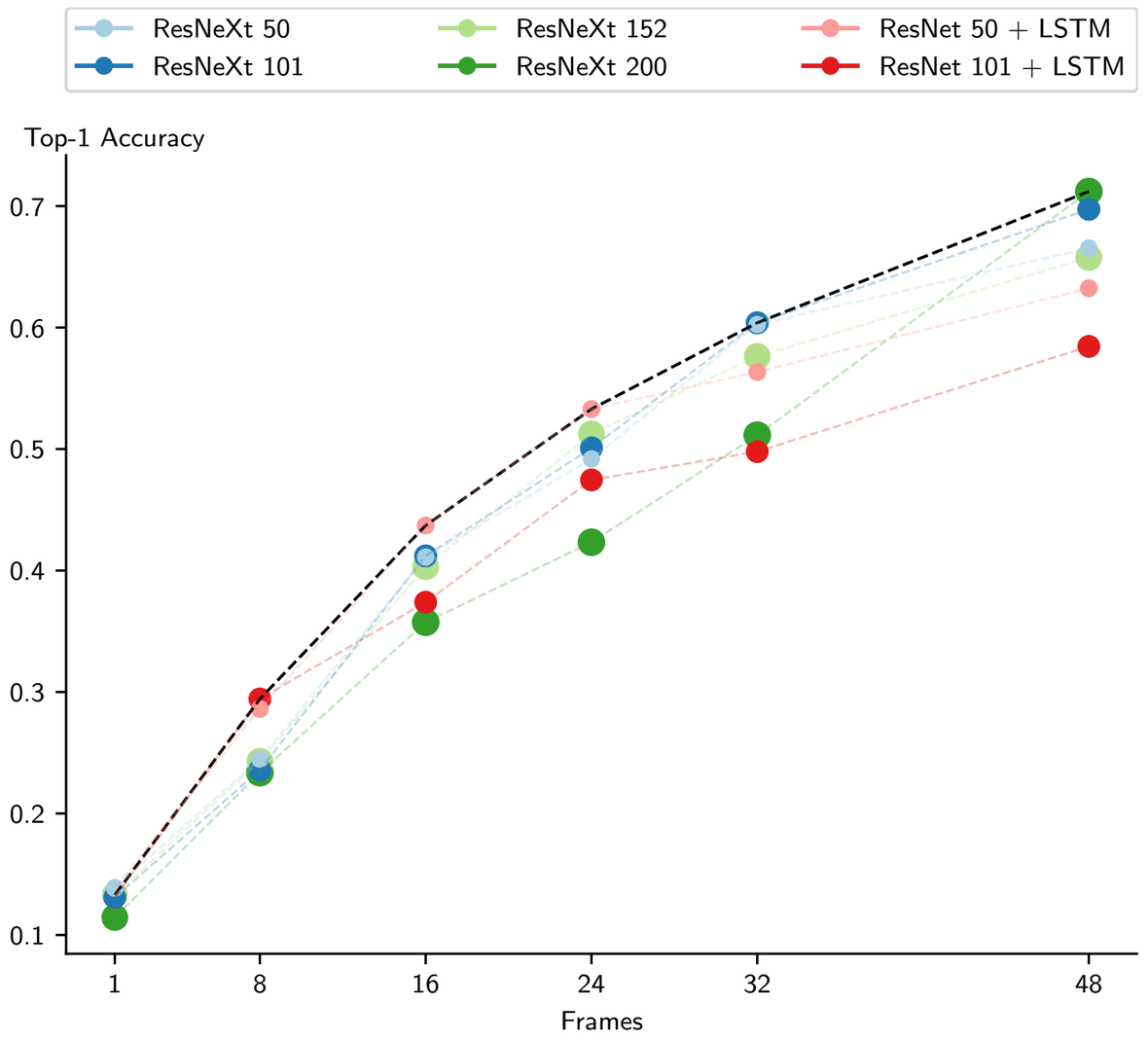


Figure 2. **Scaling Training Frames.** Performance of models across different numbers of training frames. The Pareto Frontier is represented by a black curve (—●—).

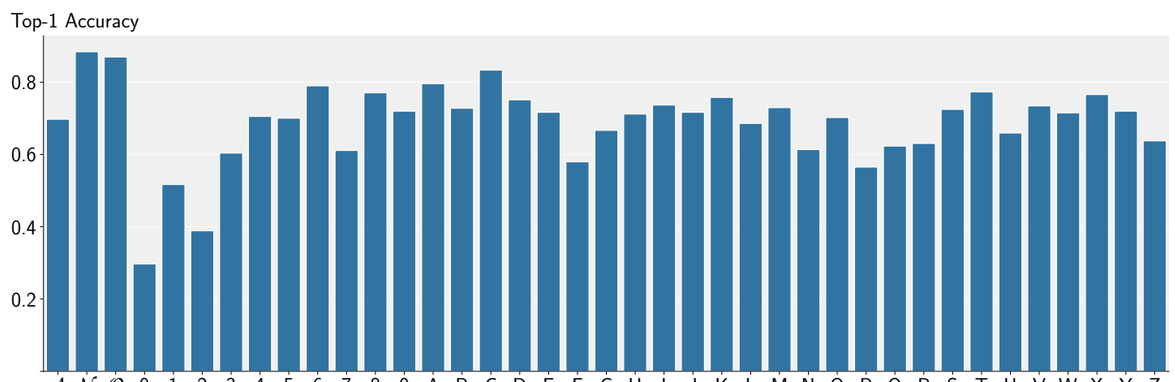


Figure 3. **Top-1 accuracy for each class** for the best-performing model from Table 2, where $\mathcal{A}$ represents the average top-1 accuracy, $\mathcal{N}$ the class "Doing Nothing" and $\mathcal{O}$ the class "Doing Other Things".

Summary of Results

- Table 2 highlights a significant gap in current end-to-end video understanding and activity recognition methods: all models, especially large vision language models, perform well below human evaluation results. Human evaluation achieves near-perfect accuracy, while the task is challenging for all tested models.
- This dataset requires models to attend through the entire video to perform well, and increasing the number of frames that models attend to significantly increases their performance (Figure 2).
- Figure 3 shows that classes such as the digits "0", "1", and "2" are particularly challenging, as they are easily confused with each other. In contrast, the contrast classes "Doing Nothing" and "Doing Other Things", are more easily recognized.