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

Rishit Dagli, Guillaume Berger, Joanna Materzynska, Ingo Bax, and Roland Memisevic

Qualcomm AI Research*
University of Toronto
MIT
rishit@cs.toronto.edu, guilberg@qti.qualcomm.com, jomat@mit.edu, {ibax, rmemisev}@qti.qualcomm.com

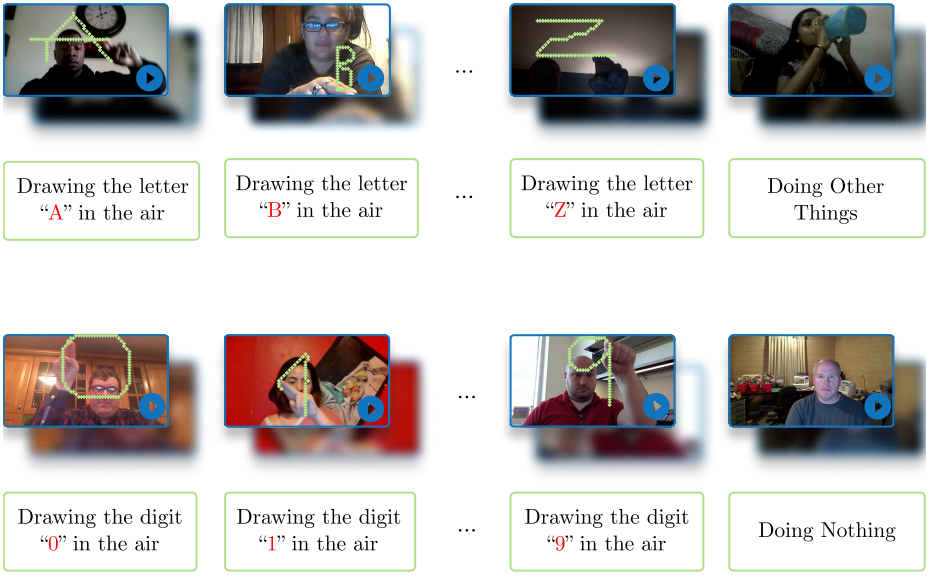

Fig. 1: Overview. We present AirLetters, a novel dataset comprised of video-label pairs of human hands denoting *characters* in the *air*. Our dataset contains videos denoting all the Latin *letters* and *digits* as well as two background classes, "Doing Other Things" and "Doing Nothing". Our dataset contains 161652 videos recorded by 1781 workers. We show the trajectory of the fingertips for visualization purposes.

Abstract. We introduce AirLetters, a new video dataset consisting of real-world videos of human-generated, articulated motions. Specifically, our dataset requires a vision model to predict letters that humans draw in the air. Unlike existing video datasets, accurate classification predictions for AirLetters rely critically on discerning motion patterns and on integrating long-range information in the video over time. An extensive

* Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

evaluation of state-of-the-art image and video understanding models on AirLetters shows that these methods perform poorly and fall far behind a human baseline. Our work shows that, despite recent progress in end-to-end video understanding, accurate representations of complex articulated motions – a task that is trivial for humans – remains an open problem for end-to-end learning.

1 Introduction

Video understanding is a long-standing research goal in AI. What makes video understanding significantly more challenging than still image understanding is that videos encode information not only spatially but also temporally, in the form of inter-frame correlations, specifically motion. Although the low-level mechanics of extracting spatio-temporal patterns from video are similar to those for extracting spatial patterns from still images (see, for example, [2]), motion understanding relies on visual features that extend across both space and time, and therefore requires operations like 3D convolutions that learn to appropriately aggregate information across these dimensions.

Real-world motion patterns, that spatio-temporal features learn to represent, range in complexity from simple transformations, for example, due to camera tilt, to spatio-temporal "textures", such as ocean waves and leaves shaking in the wind, to highly complex motion patterns that are generated by articulated (living) bodies. Although existing action recognition datasets (such as, HMDB-51 [55], UCF-101 [89], ActivityNet-200 [9], Kinetics [109], Charades [87], TikTokActions [75], as well as many others) contain patterns across this spectrum, their labels depend mostly on simple (across-frame) motion patterns and individual-frame image features. For example, to infer a label such as "Baking cookies" [9] with high confidence it suffices to look at a single frame in the video. As a result, existing video datasets make it hard to learn and evaluate a model's ability to learn complex real-world motion patterns.

An exception to this is existing datasets that focus on specific, use-case specific human-generated motion patterns. These include, in particular, datasets involving hands, which can be divided further into video sign language datasets [10, 25, 28, 30, 54, 65, 83] and general hand activity datasets [33]. However, since these datasets have been introduced with the task-specific goal of understanding sign language, gestures, or hand-object interactions, they contain a limited range of motion patterns, have already saturated performance, and in many cases also allow for inference from individual frames.

In this work, we introduce **AirLetters**, a novel dataset comprising 161652 labeled videos that capture human hand movements corresponding to digits and letters from the Latin alphabet. Our dataset is not only more challenging than existing hand gesture datasets but it also requires models to learn to precisely track hands and analyze long-term dependencies. All labels are dynamic and cannot be inferred with one or a few key frames of the video. An overview of our dataset is presented in Figure 1. The dataset also contains two

"contrast classes" labeled: "Doing Nothing" and "Doing Other Things", featuring videos of individuals engaged in tasks unrelated to positive labels. Due to the in-the-wild nature of the recordings, the videos exhibit considerable variation in lighting conditions, hand positions, backgrounds, drawing motions, and other body movements. These variations render activity recognition within our dataset particularly challenging, necessitating meticulous frame-by-frame analysis. Temporal ambiguities (for example, distinguishing between "O" and "Q" or "1" and "7") require integration over many frames. Additionally, some common types of data augmentation, such as rotation, are impractical. For example, renderings of the letters "W" and "M" appear similar under rotation. Together these challenges make our dataset a rigorous new testbed for training machine learning models to understand motion in video.

To showcase the unique challenges and opportunities our dataset presents, we conduct a series of experiments. Through these, we illustrate how our dataset supports the development of models for conventional video understanding and activity recognition. Moreover, the diversity and complexity of the video content in AirLetters makes the dataset useful both as a pre-training dataset and a benchmark in applications in which understanding the motions of human hands is important. We also hope that models focusing on video understanding or activity recognition from human hands as well as generative models that focus on generating human hands among others could benefit directly from this dataset[1].

2 Related Works

Although our data set primarily serves the purpose of learning and evaluating articulated motion understanding, it is similar in spirit to gesture and sign language recognition tasks. In this section, we provide a brief overview of existing video sign language benchmarks (§ 2.1) and video hand gesture benchmarks (§ 2.2). We also provide a brief overview of existing general video activity recognition datasets (§ 2.3).

2.1 Sign Language Datasets

Historically, the field of video sign language translation has been based on synthetic animation-based methods [18, 50, 66, 67, 81], however, such methods have been replaced by learned approaches [8, 10, 12, 17, 20, 34, 46, 53, 54] that require high-quality large-scale data.

A common way to collect sign language data involves crowd-sourcing such the videos and annotations. Many of these datasets contain comprehensive annotations for each gesture in the sign language. Widely used datasets include CSL-Daily [108] and DEVISIGN [15] in Chinese Sign Language; KETI [53] in Korean Sign Language; the Public DGS Corpus [39] in German Sign Language; LSA64 [78] in Argentinian Sign Language; PSL Kinect 30 [47] and PSL ToF [47]

[1] We plan to make data and code available at `developer.qualcomm.com`.

in Polish Sign Language; GSL [26] in Greek Sign Language; and LSE-sign [38] in Spanish Sign Language. These benchmarks feature phrases and dialogues. General word-level American Sign Language datasets include CUNY ASL [64], ASL Lexicon [4], Purdue RVL-SLLL ASL [100], and RWTH-BOSTON-50 [105], which contain general ASL words but with minimal variance among videos. Other datasets collected this way include How2Sign [25], which features instructional content translated into ASL, and sentence-level datasets like RWTH-BOSTON-104 [105] and RWTH-BOSTON-400 [105].

Some other large-scale datasets have been taken from television programs with sign language interpreters. These are often limited in variance between videos and usually have some problems with the text alignment. They include datasets like RWTH-PHOENIX-Weather [11] and SWISSTXT [13] which include weather programs in German Sign Language and Swiss German Sign Language, respectively. Other datasets derived from television programs include VRT [13] with news programs in Flemish Sign Language, and BOBSL [3] with BBC programs in British Sign Language.

Furthermore, there have also been datasets that are built by scraping videos from the web. Multiple datasets of American Sign Language have been scraped from YouTube like OpenASL [83], and YouTube-ASL [94]. There have also been datasets that scrape specialized websites, such as SP-10 [103] which includes a multilingual sign language dictionary, and AfriSign [36] which translates passages from the Bible, and The Greek Elementary School Dataset [96] with content translated from Greek elementary school content. Lastly, there are many large-scale datasets that scrape videos from the Web, containing: American Sign Language annotations: MS-ASL [45], WLASL [59], ChicagoFSWild [84], ChicagoFSWild+ [85], CISLR [44], and Indian-SL [82], all of which are word-level datasets; non-ASL annotations: SignsWorld Atlas [86], LSFB-CONT [29], LSFB-ISOL [29], ASL Fingerspelling A [74], ASL Fingerspelling B [74], PSL Fingerspelling ToF, Japanese Fingerspelling [71], RTWH Fingerspelling [24], and SIGNUM [95]; and multilingual annotations: Prompt2Sign [28].

2.2 Hand Gesture Datasets

The development of datasets in gesture recognition is primarily oriented towards enhancing the precision and versatility of gesture-based interactions in various domains, including human-computer interaction and driving assistance. This includes the Cambridge Hand Gesture dataset [51], which contains 900 RGB sequences across 9 gesture classes, and the Sheffield KInect Gesture (SKIG) dataset [62], which comprises 1080 RGB-D videos that depict dynamic gestures of 6 participants, categorizing 10 different gestures. In parallel, the ChaLearn Gesture Challenge [27, 97] contributed the ChaLearn LAP IsoGD and ConGD datasets [97], as well as the Multimodal Gesture Dataset (MMGD) [27]. Some datasets have been captured with sensors, including: MSRGesture3D 2012 [57], ChAirGest 2013 [79], Kinect Numbers and Letters Hand Gestures [76], and LTTM Senz3D [68]. Some datasets have been captured with imaging equipment. These include Interactive Museum 2014 [6], IPN Hands [7], LD-ConGR [61],

NUS HandPostures [56], FHANDS [31]. However, most of these datasets are of very small scale or do not have much variance. Exceptions are Something-Something [33] and Jester [65], which are large-scale datasets. In contrast to these, our benchmark focuses particularly on specific kinds of gestures that represent Latin characters. Another popular large-scale image dataset is the BIGHands [104] hand pose dataset, which shows significant variance between hand poses but does not represent gestures.

Multiple fine-grained datasets exist for the task of hand gesture recognition. In particular, in the context of automotive applications, datasets such as CVRR-HAND 3D [72] and nvGesture [70] are specifically designed to understand driver behavior through hand gestures, providing a controlled environment for studying gesture recognition in driving scenarios. Other specialized datasets include GUN-71 [77], which focuses on fine-grained hand movements for object manipulation, and the NATOPS [88] dataset, which focuses on air signaling gestures for airplanes.

For first-person perspective applications, datasets such as EgoHands [5], EgoFinger [43], and EgoGesture [107] offer detailed annotations for hand detection and segmentation, capturing data through wearable devices like Google Glass. This perspective is targeted at personal device interactions, and it has been extended to various specialized domains [19, 48].

2.3 Activity Recognition and Video Classification Datasets

Video classification and activity recognition involve the categorization of video content into predefined classes. UCF101 [89] consists of 13,320 video clips in 101 categories from YouTube, offering diverse and complex activities. HMDB51 [55] includes 6,766 video clips across 51 action categories from varied sources like movies and YouTube, presenting challenges such as varying camera angles and lighting. Despite its smaller size, the KTH [80] dataset, with 2,391 video sequences of six actions, laid much of the foundation of early activity recognition research. The Sports-1M [49] dataset and the Kinetics [109] series (Kinetics-400, 600, and 700) are large-scale datasets that have been instrumental in training neural networks for activity recognition tasks.

Several datasets focus on the fine-grained and contextual understanding of video content. The Charades [87] dataset, for instance, focuses on multi-label action recognition through its collection of 9,848 videos depicting everyday indoor activities across 157 action classes. They reflect real-world scenarios where multiple actions coexist. The AVA [35] dataset improves fine-grained action recognition by annotating detailed actions within 15-minute movie clips, aiding in spatiotemporal localization. Hollywood2 [58] focuses on actions in realistic settings with videos categorized into 12 human action classes and is used extensively for contextual action recognition. The COIN [91] dataset, designed for instructional video analysis, includes 11,827 videos covering 180 tasks in various domains, making it useful for understanding and segmenting instructional content. VideoLT [106] tackles the long-tailed distribution problem with its 256,218 untrimmed videos annotated in 1,004 classes, ideal for studying class imbalance.

The YouTube-8M [1] dataset, comprising 8 million videos annotated with 4,000 visual entities, serves as a large-scale benchmark for video classification models. HVU [22] aims to holistically understand videos with 572,000 videos that feature 9 million annotations on 3,142 labels.

3 The AirLetters Dataset

We present the AirLetters dataset, which is composed of short labeled videos showing people drawing letters in the air with their hands. We next provide details about our video and annotation collection method (§ 3.3), the content of the dataset (§ 3.1), and statistics of the dataset (§ 3.2).

3.1 Dataset Content

The goal of the AirLetters dataset is to provide a simple, classification-based evaluation of a model's ability to correctly understand articulated motions. 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. The dataset also includes 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, such as sitting or standing still, to represent periods of inactivity within human-computer interactions, and the "Doing Other Things" class consists of clips capturing miscellaneous, non-communicative movements such as adjusting position or random hand movements.

We show a few examples from our dataset in Figure 2 using a few frames per video. We also demonstrate the diversity of examples in our dataset in Figure 4. Our dataset is curated to reflect real-world complexity, encompassing a range of scenarios where backgrounds are often cluttered and lighting conditions vary from dimly lit to overexposed environments. This heterogeneity poses a significant challenge to the robustness of models as they have to deal with a wide spectrum of real-world conditions.

Figure 3 highlights some aspects of our dataset that are challenging for learned models but simple for humans. It shows the variability in how participants draw characters, leading to significant variation even within class. For example, the letter "B" and the digit "3" can appear quite distinct depending on the drawing styles of the participants. To accurately differentiate between these two classes, it is essential to analyze the depth and velocity of the relative motion in the videos. This analysis helps determine whether the participant intended to draw a vertical line, indicative of a "B", or merely positioned their hands, suggesting a "3". Furthermore, we also show the substantial variation in how the letter "Y" is drawn. In some cases, only the final few frames of the drawing process reveal a stroke that is crucial to differentiate "Y" from "X".

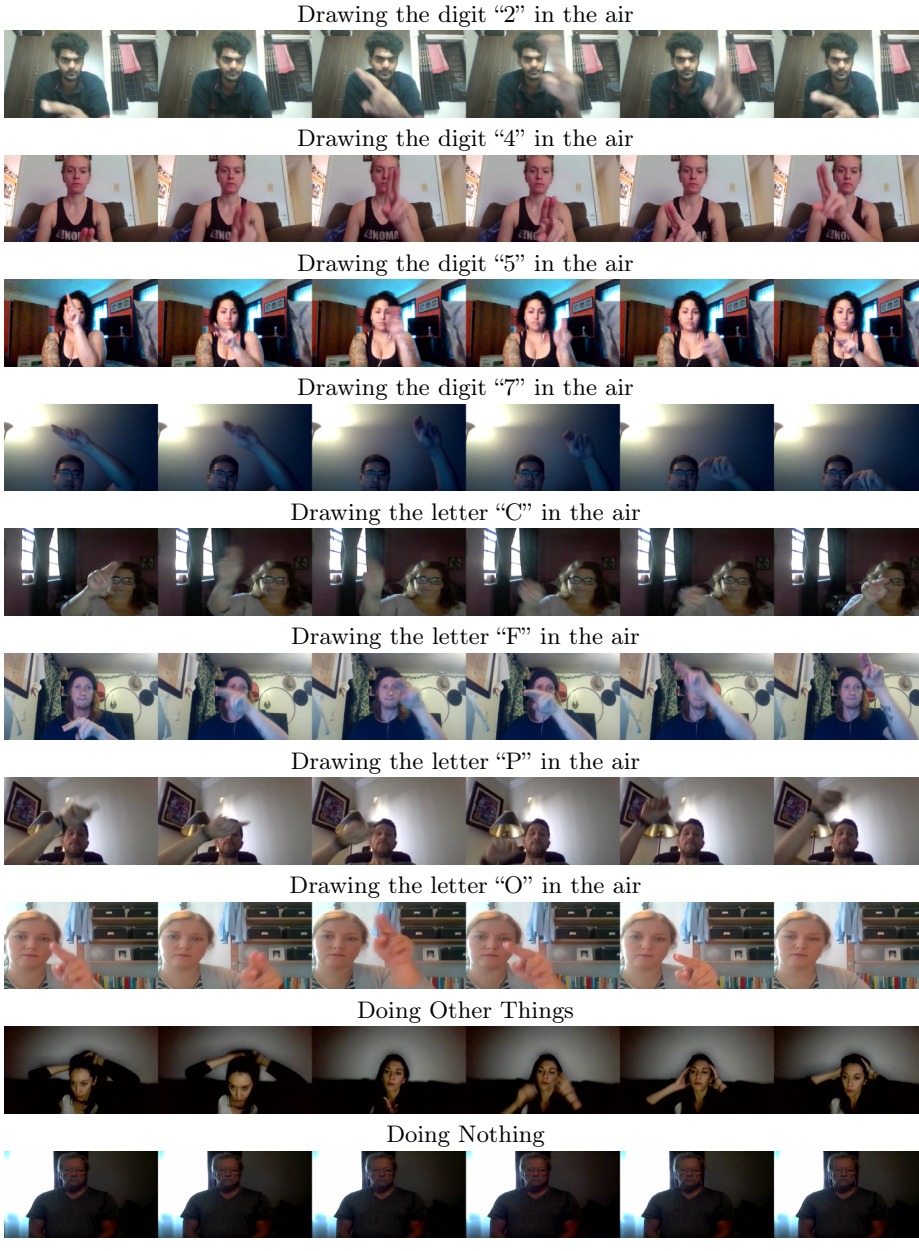

Fig. 2: Example Videos. Frames from randomly sampled videos from our dataset showing humans drawing characters as well as contrast classes.

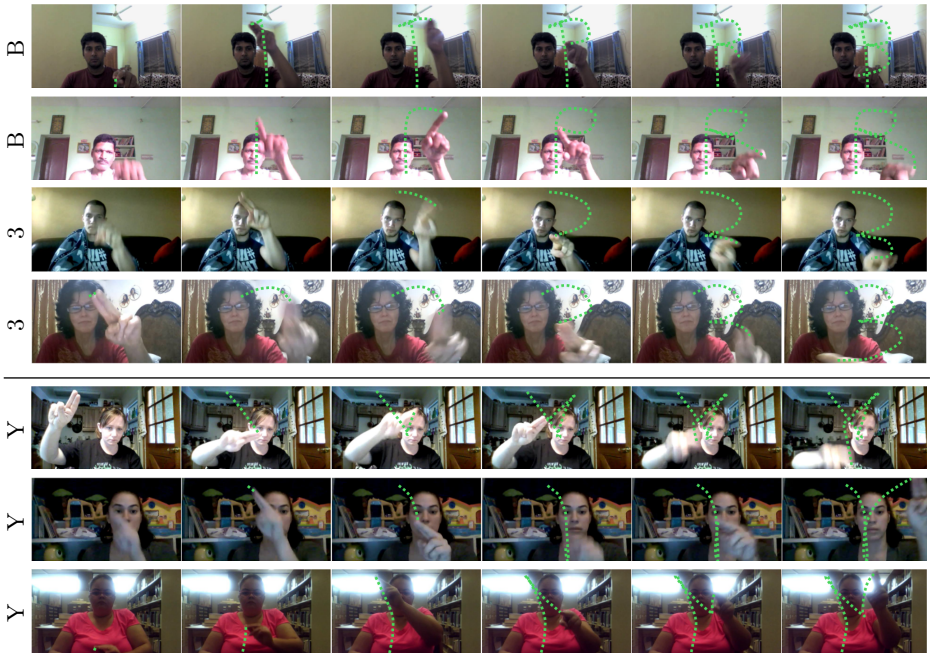

Fig. 3: Challenges due to inter-class similarities and intra-class diversity. We show some examples of drawing the letter "B" and the digit of "3", where differentiating both of these classes also requires understanding depth and velocity of relative motion to understand if the individual intended to draw a vertical line (for "B") or only meant to place their hands in position (for "3"). Underneath, we show examples of variability in drawing the letter "Y". For example, in one way version of drawing the letter "Y", only the last few frames show a stroke that distinguishes it from the letter "X".

We roughly split the dataset using an 8:1:1 ratio for training, validation, and testing, respectively. To do so, we assign each one of the 1781 crowd workers to either the training, validation, or test split. We show the number of videos in each split in Table 1.

Table 1: Dataset Splits. The number of crowd workers and videos in each split of our dataset.

Split	Videos	Workers
Train	128745	958
Validation	16480	412
Test	16427	411

3.2 Dataset Statistics

Our dataset is designed to mirror real-world conditions and showcases a diverse range of backgrounds and variations, with 1781 crowd workers contributing. It consists of 161652 videos, with each contributor recording an average of 90.76 videos at an average frame rate of 30 frames per second (fps) to accommodate different recording devices. The average duration of each video is approximately 2.92 seconds, allowing for the completion of the required gesture without pro-

Fig. 4: 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.

longing the recording unnecessarily. The spatial resolution of the videos has an area averaging 0.25 megapixels at varying aspect ratios. The total number of frames per video varies depending on the frame rate, but on average, each video contains ≈252 frames. We show summaries of these statistics in Table 2.

3.3 Collection Methodology

To collect our dataset, we used a custom platform integrated with crowd-sourcing providers. This allowed us to recruit participants from diverse gender, geographical, and ethnic backgrounds and to provide the required instruction and recording functionality. Participants redirected to our platform were asked to record themselves performing all 36 gestures in front of their camera. We provided detailed visual and textual instructions to ensure clear hand visibility, high video quality, and precise gesture execution. Supplementary example videos were provided to demonstrate correct gestures and to address the limitations of text instructions. After reviewing the guidelines, participants prepared for recording with the help of a countdown timer. Recordings averaged ≈3 seconds, after which participants could review and re-record if necessary. For added variability, the "Doing Other Things" category required four distinct activities, while the

Table 2: 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)

"Doing Nothing" category required no specific activities. Each participant could make up to three submissions. To encourage scene variability, participants could interrupt recording and resume at a later time.

All submissions were reviewed by human operators to verify accuracy. Participants with mostly correct submissions but minor errors were allowed to make corrections and resubmit. This approach ensured the high quality and consistency of the dataset. Finally, all videos were resized to a width of 640 pixels, maintaining the aspect ratio.

4 Experiments Validating AirLetters

We conduct various experiments to assess the difficulty of this task. Below, we highlight the baseline architectures we used (§ 4.1), our preprocessing workflow (§ 4.2), and present our results (§ 4.3).

4.1 Baseline Architectures

Image Models. We train baseline image classification models, including ResNet [41], ResNeXt [102], SE ResNeXt [98], MaxViT [93], and ViT [23] to predict the activity label given a single video frame. During testing, we average model outputs for each frame of the test videos to produce a final prediction.

Video Models. We also train baseline video models, including ResNet 3D, ResNeXt 3D, Strided Inflated EfficientNet 3D [69], and VideoMAE [92]. Since the data are inherently temporal, we also train a ResNet [41] baseline paired with an LSTM [42], where the ResNet backbone extracts 2D features from individual frames and the features are passed to an LSTM layer. We use the last hidden state as the encoding for the videos. We compare training from scratch, finetuning from models pre-trained on either Kinetics [109] or ImageNet [21], as well as finetuning from Imagenet pre-trained classifiers whose parameters are inflated to 3D [14].

Table 3: Classification accuracy of multiple image models, video models, and (large) vision language models on the AirLetters dataset. Note that the task is straightforward for humans but challenging for existing models.

Method	**Top-1 Accuracy (↑)**
Image Models	
ViT-B/16 [23]	7.49
MaxViT-T [93]	7.56
ResNet-200 [41]	11.44
ResNeXt-101 [102]	13.09
SE-ResNeXt-26 [98]	13.29
ResNet-50 [41]	13.87
Video Models	
VideoMAE (16) [92]	57.96
ResNet-101 + LSTM	58.45
ResNet-50 + LSTM	63.24
ResNext-152 3D	65.77
Strided Inflated EfficientNet 3D [69]	65.97
ResNext-50 3D	66.54
ResNext-101 3D	69.74
ResNext-200 3D	71.20
Vision Language Models	
Video-LLaVA (w/o contrast class) [60]	2.53
VideoLLaMA2 (w/o contrast class) [16]	2.47
Video-LLaVA [60]	7.29
VideoLLaMA2 [16]	7.58
Human Performance (10 videos/class)	96.67

Vision Language Models. We also experiment with identifying actions from videos in a zero-shot manner using a large vision language model, specifically Video LLaVa [60] and Video Llama2 [16]. This includes experiments, where we remove the two contrast classes we have while evaluating these models to demonstrate the difficulty they have in estimating the non-contrast classes.

We train the baseline models using either Adam [52] or AdamW [63] and adopt the standard cross-entropy loss with label smoothing [90]. Depending on the model, we experiment with various learning rates and schedules, including constant learning rates, cosine decay, and exponential decay. We present experimental details in Appendix A.

4.2 Preprocessing Workflow

Before training, we resample all of our videos to 30 FPS and resize the videos in an aspect-ratio preserving manner to 300 pixels for the shortest edge. We use a standard video processing workflow during training and testing. We sample the

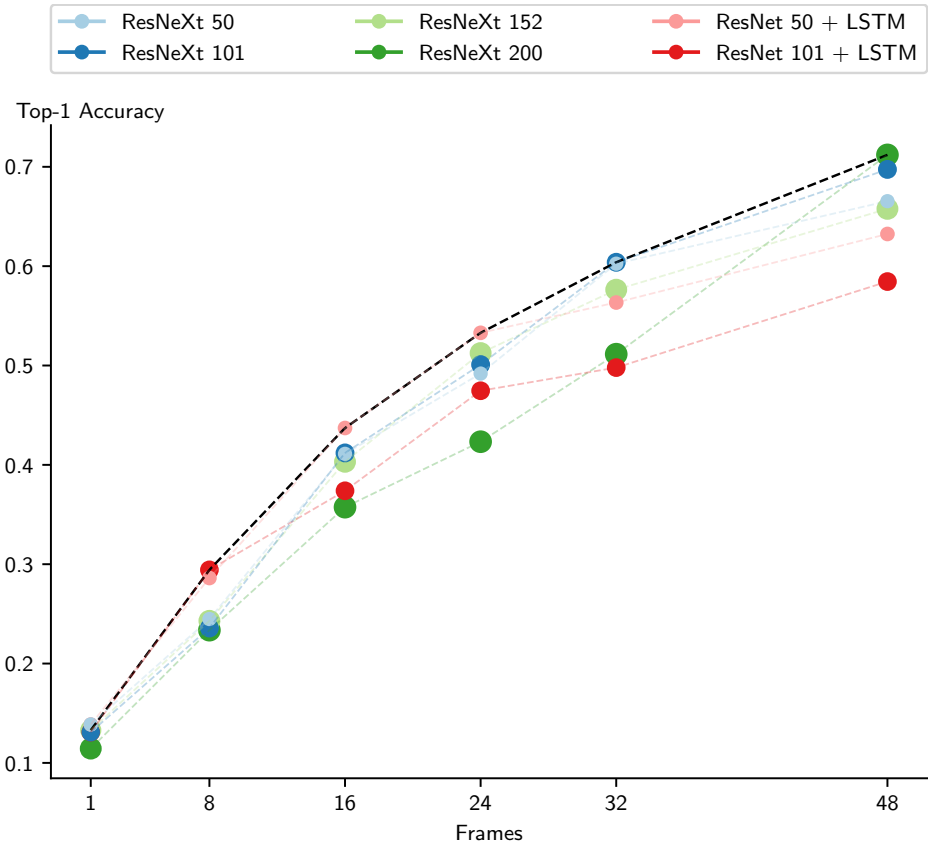

Fig. 5: Scaling Training Frames. Performance of models across different numbers of training frames. The Pareto Frontier is represented by a black curve (—●—). Note that this dataset requires models to attend through the entire video to perform well, and increasing the number of frames that models attend to significantly increases their performance.

videos with FPS $\in [8, 24]$. In the case that sampling at 8 FPS does not leave us with at least the number of frames required for the model, we shift the lower bound to an FPS that can give us at least the number of frames needed. We then perform a spatio-temporal crop on the videos. During evaluation and testing, we perform a center crop, followed by sampling the required number of frames by performing a temporal center crop.

4.3 Results

We evaluated the baseline architectures described in Section § 4.1 for the task of end-to-end video activity recognition on our data set, and report the top-1 accuracy in Table 3. Our results highlight a significant gap in current end-to-

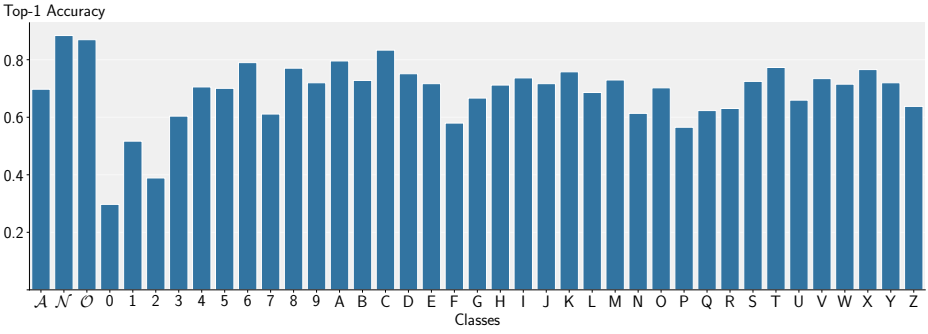

Fig. 6: Top-1 accuracy for each class for the best-performing model from Table 3, where $\mathcal{A}$ represents the average top-1 accuracy, $\mathcal{N}$ the class "Doing Nothing" and $\mathcal{O}$ the class "Doing Other Things".

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.

We examine the performance of models trained on different numbers of frames: 1 (image models), 8, 16, 24, 32, and 48 frames per video in Figure 5. We observe a significant increase in the performance of models when they are trained on more frames. On average, the videos are sampled at 16 FPS (due to our pre-processing described in Section 4.2) and have a duration of approximately seconds (Section 3). We notice a significant increase in performance when increasing the number of frames from 32 to 48, demonstrating that our dataset requires models to attend to most frames of the video to perform well on this benchmark. Furthermore, our experiments also validate that our dataset requires models to learn long-range temporal dependencies and to have the ability to aggregate information temporally.

We also show the top-1 accuracy for each of the classes of a ResNeXt-3D model in Figure 6 and the corresponding confusion matrix in Figure 7. We observe 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. We also notice some expected misclassification patterns in Figure 7, such as "0" and "O", "3" and "B", or "P" and "D" being misclassified for one another due to the visual similarity of these characters.

5 Conclusion

We introduced a new real-world dataset, utilizing human generated articulated motions. Unlike existing video datasets, accurate predictions for our dataset require detailed understanding of motion and the integration of long-range information across the video. We show that existing image and video understanding

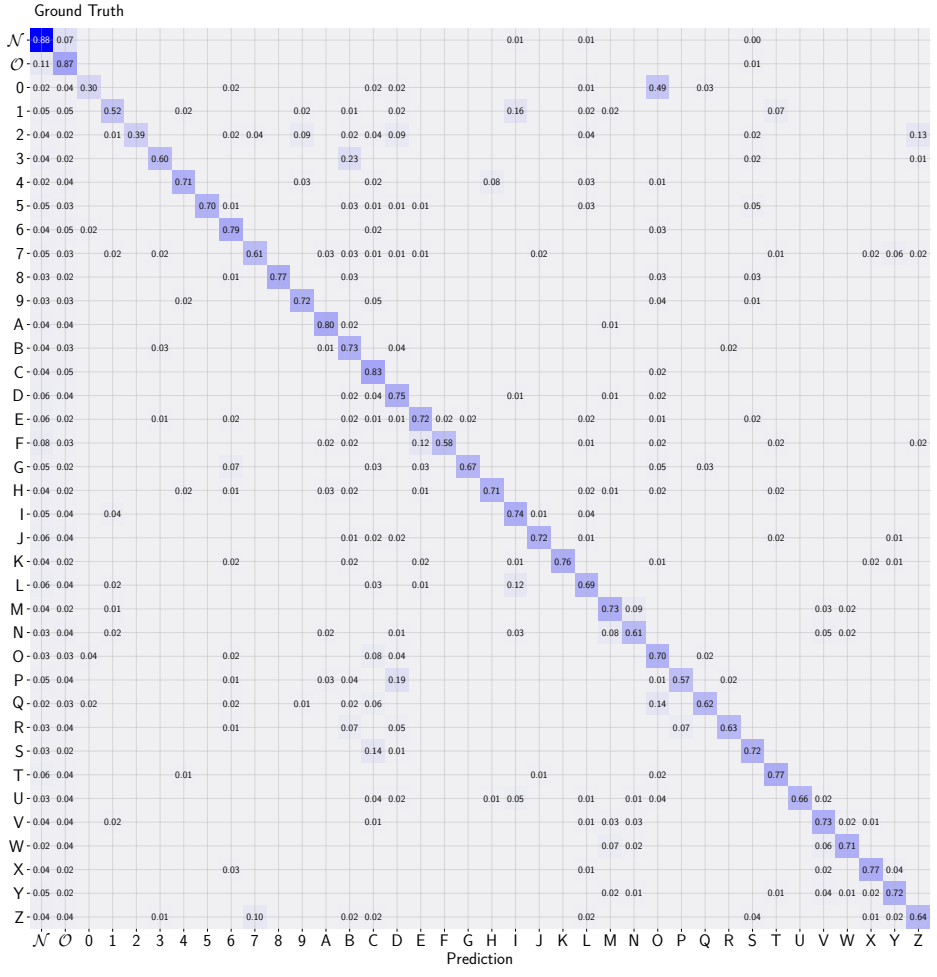

Fig. 7: Confusion Matrix for the best-performing model from Table 3, where $\mathcal{N}$ represents the class "Doing Nothing" and $\mathcal{O}$ represents the class "Doing Other Things".

models perform poorly and fall far behind the human baseline. In this way, our work exposes a significant gap in the current video understanding capabilities. Closing this gap, arguably, will be a necessary step to build AI models that can perceive the world more like humans.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A.P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. In: arXiv:1609.08675 (2016), https://arxiv.org/pdf/1609.08675v1.pdf
2. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America. A, Optics and image science **2 2**, 284–99 (1985), https://api.semanticscholar.org/CorpusID:5248006
3. Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., Zisserman, A.: Bbc-oxford british sign language dataset (2021)
4. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A.: The american sign language lexicon video dataset. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8 (2008). https://doi.org/10.1109/CVPRW.2008.4563181
5. Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
6. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2014)
7. Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., Yanai, K.: Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In: 25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, Jan 10–15, 2021. pp. 4340–4347. IEEE (2021)
8. Boháček, M., Hrúz, M.: Sign pose-based transformer for word-level sign language recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 182–191 (January 2022)
9. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
10. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
11. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7784–7793 (2018). https://doi.org/10.1109/CVPR.2018.00812
12. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
13. Camgoz, N.C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., Bowden, R.: Content4all open research sign language translation datasets (2021)
14. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (2017). https://doi.org/10.1109/CVPR.2017.502

15. Chai, X., Wanga, H., Zhoub, M., Wub, G., Lic, H., Chena, X.: Devisign: dataset and evaluation for 3d sign language recognition. Technical report, Beijing, Tech. Rep (2015)
16. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms (2024), https://arxiv.org/abs/2406.07476
17. Cooper, H., Bowden, R.: Large lexicon detection of sign language. In: Human–Computer Interaction: IEEE International Workshop, HCI 2007 Rio de Janeiro, Brazil, October 20, 2007 Proceedings 4. pp. 88–97. Springer (2007)
18. Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., Abbott, S.: Tessa, a system to aid communication with deaf people. In: Proceedings of the Fifth International ACM Conference on Assistive Technologies. p. 205–212. Assets '02, Association for Computing Machinery, New York, NY, USA (2002). https://doi.org/10.1145/638249.638287, https://doi.org/10.1145/638249.638287
19. Crispim-Junior, C.F., Buso, V., Avgerinakis, K., Meditskos, G., Briassouli, A., Benois-Pineau, J., Kompatsiaris, I.Y., Bremond, F.: Semantic event fusion of different visual modality concepts for activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1598–1611 (2016). https://doi.org/10.1109/TPAMI.2016.2537323
20. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
21. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
22. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 593–610. Springer International Publishing, Cham (2020)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
24. Dreuw, P., Deselaers, T., Keysers, D., Ney, H.: Modeling image variability in appearance-based gesture recognition. In: ECCV workshop on statistical methods in multi-image and video processing. pp. 7–18 (2006)
25. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
26. Efthimiou, E., Fotinea, S.E.: Gslc: Creation and annotation of a greek sign language corpus for hci. In: Stephanidis, C. (ed.) Universal Acess in Human Computer Interaction. Coping with Diversity. pp. 657–666. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
27. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Sigal, L., Argyros, A., Sminchisescu, C., Bowden, R., Sclaroff, S.: Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In: Proceedings of the 15th ACM on International Conference on Multimodal

Interaction. p. 365–368. ICMI '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2522848.2532597, https://doi.org/10.1145/2522848.2532597
28. Fang, S., Wang, L., Zheng, C., Tian, Y., Chen, C.: Signllm: Sign languages production large language models (2024)
29. Fink, J., Frénay, B., Meurant, L., Cleve, A.: Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2021). https://doi.org/10.1109/IJCNN52387.2021.9534336
30. Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J.H., Ney, H.: Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In: LREC. vol. 9, pp. 3785–3789 (2012)
31. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
32. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
33. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
34. Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. vol. 1, pp. 162–167 vol.1 (1997). https://doi.org/10.1109/ICSMC.1997.625742
35. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
36. Gueuwou, S., Takyi, K., Müller, M., Nyarko, M.S., Adade, R., Gyening, R.M.O.M.: Afrisign: Machine translation for african sign languages. In: 4th Workshop on African Natural Language Processing (2023), https://openreview.net/forum?id=EHldk3J2xk
37. Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., Sun, M., Bossan, B.: Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate (2022)
38. Gutierrez-Sigut, E., Costello, B., Baus, C., Carreiras, M.: Lse-sign: A lexical database for spanish sign language. Behavior Research Methods **48**, 123–137 (2016)
39. Hanke, T., Schulder, M., Konrad, R., Jahn, E.: Extending the public dgs corpus in size and depth. In: sign-lang@ LREC 2020. pp. 75–82. European Language Resources Association (ELRA) (2020)
40. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv:2111.06377 (2021)
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385

42. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
43. Huang, Y., Liu, X., Zhang, X., Jin, L.: A pointing gesture based egocentric interaction system: Dataset, approach and application. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2016)
44. Joshi, A., Bhat, A., Pradeep, S., Gole, P., Gupta, S., Agarwal, S., Modi, A.: Cislr: Corpus for indian sign language recognition. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10357–10366 (2022)
45. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language (2019)
46. Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: BMVC. pp. 1–10 (2004)
47. Kapuscinski, T., Oszust, M., Wysocki, M., Warchol, D.: Recognition of hand gestures observed by depth cameras. International Journal of Advanced Robotic Systems **12**(4), 36 (2015). https://doi.org/10.5772/60091, https://doi.org/10.5772/60091
48. Karaman, S., Benois-Pineau, J., Dovgalecs, V., Mégret, R., Pinquier, J., André-Obrecht, R., Gaëstel, Y., Dartigues, J.F.: Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia. Multimedia tools and applications **69**, 743–771 (2014)
49. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
50. Karpouzis, K., Caridakis, G., Fotinea, S.E., Efthimiou, E.: Educational resources and implementation of a greek sign language synthesis architecture. Computers & Education **49**(1), 54–74 (2007). https://doi.org/https://doi.org/10.1016/j.compedu.2005.06.004, https://www.sciencedirect.com/science/article/pii/S0360131505000849, web3D Technologies in Learning, Education and Training
51. Kim, T.K., Cipolla, R.: Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(8), 1415–1428 (2009). https://doi.org/10.1109/TPAMI.2008.167
52. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), https://arxiv.org/abs/1412.6980
53. Ko, S.K., Kim, C.J., Jung, H., Cho, C.: Neural sign language translation based on human keypoint estimation. Applied sciences **9**(13), 2683 (2019)
54. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015)
55. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563 (2011). https://doi.org/10.1109/ICCV.2011.6126543
56. Kumar, P., Vadakkepat, P., Loh, A.: Hand posture and face recognition using a fuzzy-rough approach (2010)

57. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). pp. 1975–1979 (2012)
58. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587756
59. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1448–1458 (2020). https://doi.org/10.1109/WACV45572.2020.9093512
60. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2023)
61. Liu, D., Zhang, L., Wu, Y.: Ld-congr: A large rgb-d video dataset for long-distance continuous gesture recognition. In: CVPR (2022)
62. Liu, L., Shao, L.: Learning discriminative representations from rgb-d video data. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. p. 1493–1500. IJCAI '13, AAAI Press (2013)
63. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
64. Lu, P., Huenerfauth, M.: Collecting and evaluating the cuny asl corpus for research on american sign language animation. Computer Speech & Language **28**(3), 812–831 (2014). https://doi.org/https://doi.org/10.1016/j.csl.2013.10.004, https://www.sciencedirect.com/science/article/pii/S0885230813000879
65. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
66. Mazumder, S., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V.: Translating sign language videos to talking faces. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing. ICVGIP '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3490035.3490286, https://doi.org/10.1145/3490035.3490286
67. McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D.G., Stumbo, M., Berke, L., Bialek, M., Thomas, F.: An automated technique for real-time production of lifelike animations of american sign language. Universal Access in the Information Society **15**, 551–566 (2016)
68. Memo, A., Zanuttigh, P.: Head-mounted gesture controlled interface for human-computer interaction. Multimedia Tools and Applications **77**, 27–53 (2018)
69. Mercier, A., Berger, G., Panchal, S., Letsch, F., Boehm, C., Kang, N., Bax, I., Memisevic, R.: Is end-to-end learning enough for fitness activity recognition? arXiv preprint arXiv:2305.08191 (2023)
70. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
71. Nguen, N.T., Sako, S., Kwolek, B.: Deep cnn-based recognition of jsl finger spelling. In: Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, 2019, Proceedings 14. pp. 602–613. Springer (2019)

72. Ohn-Bar, E., Trivedi, M.M.: Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. IEEE Transactions on Intelligent Transportation Systems **15**(6), 2368–2377 (2014). https://doi.org/10.1109/TITS.2014.2337331
73. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
74. Pugeault, N., Bowden, R.: Spelling it out: Real-time asl fingerspelling recognition. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp. 1114–1119 (2011). https://doi.org/10.1109/ICCVW.2011.6130290
75. Qian, Y., Sun, Y., Kargarandehkordi, A., Mutlu, O.C., Surabhi, S., Chen, P., Jabbar, Z., Wall, D.P., Washington, P.: Advancing human action recognition with foundation models trained on unlabeled public videos (2024)
76. Quiroga, F., Corbalán, L.C.: A novel competitive neural classifier for gesture recognition with small training sets. In: XVIII Congreso Argentino de Ciencias de la Computación (CACIC) (2013)
77. Rogez, G., Supancic, III, J.S., Ramanan, D.: Understanding everyday hands in action from rgb-d images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
78. Ronchetti, F., Quiroga, F.M., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: An argentinian sign language dataset (2023)
79. Ruffieux, S., Lalanne, D., Mugellini, E.: Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. p. 483–488. ICMI '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2522848.2532590, https://doi.org/10.1145/2522848.2532590
80. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. vol. 3, pp. 32–36 Vol.3 (2004). https://doi.org/10.1109/ICPR.2004.1334462
81. Segouat, J.: A study of sign language coarticulation. SIGACCESS Access. Comput. (93), 31–38 (jan 2009). https://doi.org/10.1145/1531930.1531935, https://doi.org/10.1145/1531930.1531935
82. Selvaraj, P., NC, G., Kumar, P., Khapra, M.: Openhands: Making sign language recognition accessible with pose-based pretrained models across languages (2021)
83. Shi, B., Brentari, D., Shakhnarovich, G., Livescu, K.: Open-domain sign language translation learned from online video (2022)
84. Shi, B., Del Rio, A.M., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., Livescu, K.: American sign language fingerspelling recognition in the wild. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 145–152 (2018). https://doi.org/10.1109/SLT.2018.8639639
85. Shi, B., Rio, A.M.D., Keane, J., Brentari, D., Shakhnarovich, G., Livescu, K.: Fingerspelling recognition in the wild with iterative visual attention. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5399–5408 (2019). https://doi.org/10.1109/ICCV.2019.00550

86. Shohieb, S.M., Elminir, H.K., Riad, A.: Signsworld atlas; a benchmark arabic sign language database. Journal of King Saud University - Computer and Information Sciences **27**(1), 68–76 (2015). https://doi.org/https://doi.org/10.1016/j.jksuci.2014.03.011, https://www.sciencedirect.com/science/article/pii/S1319157814000548
87. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 510–526. Springer International Publishing, Cham (2016)
88. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 500–506 (2011). https://doi.org/10.1109/FG.2011.5771448
89. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
90. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
91. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
92. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 10078–10093. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf
93. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. p. 459–479. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20053-3_27, https://doi.org/10.1007/978-3-031-20053-3_27
94. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus (2023)
95. Von Agris, U., Kraiss, K.F.: Towards a video corpus for signer-independent continuous sign language recognition. Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May **11**(2) (2007)
96. Voskou, A., Panousis, K.P., Partaourides, H., Tolias, K., Chatzis, S.: A new dataset for end-to-end sign language translation: The greek elementary school dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1966–1975 (October 2023)
97. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2016)
98. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
99. Wightman, R., Touvron, H., Jegou, H.: Resnet strikes back: An improved training procedure in timm. In: NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future (2021), https://openreview.net/forum?id=NG6MJnVl6M5

100. Wilbur, R., Kak, A.C.: Purdue rvl-slll american sign language database (2006)
101. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing. pp. 38–45. Association for Computational Linguistics (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
102. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
103. Yin, A., Zhao, Z., Jin, W., Zhang, M., Zeng, X., He, X.: Mlslt: Towards multilingual sign language translation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5099–5109 (2022). https://doi.org/10.1109/CVPR52688.2022.00505
104. Yuan, S., Ye, Q., Stenger, B., Jain, S., Kim, T.K.: Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
105. Zahedi, M., Keysers, D., Deselaers, T., Ney, H.: Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In: Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27. pp. 401–408. Springer (2005)
106. Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.G., Davis, L.: Videolt: Large-scale long-tailed video recognition (2021)
107. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia **20**(5), 1038–1050 (2018). https://doi.org/10.1109/TMM.2018.2808769
108. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1316–1325 (June 2021)
109. Zisserman, A., Carreira, J., Simonyan, K., Kay, W., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M.: The kinetics human action video dataset (2017)

A Experimental Details

We show the prompts we used to evaluate the vision language models in Figure 8.

All of our experiments use PyTorch 1.13 [73] and Accelerate [37] to train our models in a distributed fashion. Our codebase is based on `timm` [98] and HuggingFace Transformers [101]. We present all the experimental details for training the models from Table 3 in Tables 4 to 6, for all other hyper-parameters we use PyTorch defaults.

B Additional Examples

We demonstrate additional video examples from our dataset in Figure 9. We also show a histogram showing the distribution of videos among classes. All classes have an equal number of videos except "doing other things" which has $3\times$ the number of videos.

Table 4: Experimental Details for Video Models I, whose performance is shown in Table 3.

Model	ResNeXt-200 3D	Strided Inflated EfficientNet 3D [69]
Training Precision	FP-32	FP-32
# of frames	48	48
Frame Size	$(224, 224)$	$(224, 224)$
Initialization	ImageNet-1k w/ RA1 recipe [99]	Properiatery Dataset w/ recipe [69]
Training Preprocessing	Random Resized Crop, $(0.7, 1.0)$	Random Resized Crop, $(0.7, 1.0)$
Eval Preprocessing	Center Crop	Center Crop
Label Smoothing	10^{-1}	10^{-1}
Batch Size	32	32
Optimizer	AdamW [63]	AdamW [63]
Optimizer Parameters	$\lambda = 10^{-2}$	$\lambda = 10^{-2}$
	$\beta_1 = 0.9$	$\beta_1 = 0.9$
	$\beta_2 = 0.999$	$\beta_2 = 0.999$
Initial learning rate	10^{-4}	10^{-3}
LR Schedule	Static	Static
Scheduler Parameters		
Gradient clipping	None	None
Training Iterations	385k	385k
Params (M)	67.77	14.46

Table 5: Experimental Details for Video Models II, whose performance is shown in Table 3.

Model	VideoMAE [92]	ResNet-101 + LSTM	ResNet-50 + LSTM
Training Precision	FP-32	FP-32	FP-32
# of frames	16	48	48
Frame Size	$(224, 224)$	$(224, 224)$	$(224, 224)$
Initialization	ImageNet-1k w/ MAE [40]	ImageNet-1k w/ RA1 recipe [99]	ImageNet-1k w/ RA1 recipe [99]
Training Preprocessing	Random Resized Crop, $(0.7, 1.0)$	Random Resized Crop, $(0.7, 1.0)$	Random Resized Crop, $(0.7, 1.0)$
Eval Preprocessing	Center Crop	Center Crop	Center Crop
Label Smoothing	10^{-1}	10^{-1}	10^{-1}
Batch Size	8	8	32
Optimizer	AdamW [63]	Adam [52]	Adam [52]
Optimizer Parameters	$\lambda = 10^{-2}$	$\lambda = 0$	$\lambda = 0$
	$\beta_1 = 0.9$	$\beta_1 = 0.9$	$\beta_1 = 0.9$
	$\beta_2 = 0.999$	$\beta_2 = 0.999$	$\beta_2 = 0.999$
Initial learning rate	10^{-5}	10^{-4}	10^{-3}
LR Schedule	Cosine Annealing w/ Warm Restart	Static	Static
Scheduler Parameters	$T_0 = 2$		
	$\eta_{\min} = 10^{-2}$		
	$\alpha_{max} = 10^{-2}$		
Gradient clipping	None	None	None
Training Iterations	600k	385k	385k
Params (M)	86.26	43.72	24.73

Model	Resnext-50 3D	Resnext-101 3D	Resnext-152 3D
Training Precision	FP-32	FP-32	FP-32
# of frames	48	48	48
Frame Size	$(224, 224)$	$(224, 224)$	$(224, 224)$
Initialization	ImageNet-1k w/ RA1 recipe [99]	ImageNet-1k w/ RA1 recipe [99]	ImageNet-1k w/ RA1 recipe [99]
Training Preprocessing	Random Resized Crop, $(0.7, 1.0)$	Random Resized Crop, $(0.7, 1.0)$	Random Resized Crop, $(0.7, 1.0)$
Eval Preprocessing	Center Crop	Center Crop	Center Crop
Label Smoothing	10^{-1}	10^{-1}	10^{-1}
Batch Size	32	32	64
Optimizer	Adam [52]	Adam [52]	AdamW [63]
Optimizer Parameters	$\lambda = 0$	$\lambda = 0$	$\lambda = 10^{-2}$
	$\beta_1 = 0.9$	$\beta_1 = 0.9$	$\beta_1 = 0.9$
	$\beta_2 = 0.999$	$\beta_2 = 0.999$	$\beta_2 = 0.999$
Initial learning rate	10^{-4}	10^{-4}	10^{-4}
LR Schedule	Static	Static	Static
Scheduler Parameters			
Gradient clipping	None	None	None
Training Iterations	385k	385k	385k
Params (M)	23.17	44.82	62.66

Table 6: Experimental Details for Image Models, whose performance is shown in Table 3.

Model	ViT-B/16 [23]	MaxViT-T [93]	ResNet-200 [41]
Training Precision	FP-32	FP-32	FP-32
Frame Size	$(224, 224)$	$(224, 224)$	$(224, 224)$
Initialization	ImageNet-1k w/ MAE [40]	ImageNet-1k (TF Weights)	ImageNet-1k w/ RA2 recipe [99]
Training Preprocessing	Center Crop	Center Crop	Center Crop
Eval Preprocessing	Center Crop	Center Crop	Center Crop
Label Smoothing	10^{-1}	10^{-1}	10^{-1}
Batch Size	512	64	512
Optimizer	AdamW [63]	AdamW [63]	AdamW [63]
Optimizer Parameters	$\lambda = 5 \times 10^{-2}$	$\lambda = 5 \times 10^{-2}$	$\lambda = 5 \times 10^{-2}$
	$\beta_1 = 0.9$	$\beta_1 = 0.9$	$\beta_1 = 0.9$
	$\beta_2 = 0.999$	$\beta_2 = 0.999$	$\beta_2 = 0.999$
Initial learning rate	10^{-5}	10^{-3}	10^{-3}
LR Schedule	Cosine Annealing w/ Warm Restart	Cosine Annealing w/ Warm Restart	Cosine Annealing w/ Warm Restart
Scheduler Parameters	$T_0 = 2$	$T_0 = 2$	$T_0 = 2$
	$\eta_{\min} = 10^{-2}$	$\eta_{\min} = 10^{-2}$	$\eta_{\min} = 10^{-2}$
	$\alpha_{max} = 10^{-2}$	$\alpha_{max} = 10^{-2}$	$\alpha_{max} = 10^{-2}$
Gradient clipping	None	None	None
Training Iterations	230k	230k	180k
Params (M)	85.83	28.56	62.72

Model	ResNeXt-101 [102]	SE ResNeXt [98]	ResNet-50 [41]
Training Precision	FP-32	FP-32	FP-32
Frame Size	$(224, 224)$	$(224, 224)$	$(224, 224)$
Initialization	ImageNet-1k w/ MAE [40]	YFCC100M FT ImageNet-1k	ImageNet-1k w/ RA1 recipe [99]
Training Preprocessing	Center Crop	Center Crop	Center Crop
Eval Preprocessing	Center Crop	Center Crop	Center Crop
Label Smoothing	10^{-1}	10^{-1}	10^{-1}
Batch Size	1024	1024	1024
Optimizer	AdamW [63]	AdamW [63]	AdamW [63]
Optimizer Parameters	$\lambda = 5 \times 10^{-2}$	$\lambda = 5 \times 10^{-2}$	$\lambda = 5 \times 10^{-2}$
	$\beta_1 = 0.9$	$\beta_1 = 0.9$	$\beta_1 = 0.9$
	$\beta_2 = 0.999$	$\beta_2 = 0.999$	$\beta_2 = 0.999$
Initial learning rate	10^{-4}	10^{-4}	10^{-3}
LR Schedule	Cosine Annealing w/ Warm Restart	Cosine Annealing w/ Warm Restart	Cosine Annealing w/ Warm Restart
Scheduler Parameters	$T_0 = 2$	$T_0 = 2$	$T_0 = 2$
	$\eta_{\min} = 10^{-2}$	$\eta_{\min} = 10^{-2}$	$\eta_{\min} = 10^{-2}$
	$\alpha_{max} = 10^{-2}$	$\alpha_{max} = 10^{-2}$	$\alpha_{max} = 10^{-2}$
Gradient clipping	None	None	None
Training Iterations	180k	180k	180k
Params (M)	42.58	14.84	23.59

Video-LLaVA

Q: `USER`: `<video>`You are a video classifier trained to detect letters and digits drawn by humans in the air with their fingers. You will be provided with a video of a person drawing in the air, carefully analyze the video to determine what letter or digit the person draws. Only respond with the character detected, no explanation. The only valid responses are "doing other things" or "doing nothing" or a letter or a digit. `ASSISTANT`:

A: doing other things

Video-LLaVA (w/o contrast classes)

Q: `USER`: `<video>`You are a video classifier trained to detect letters and digits drawn by humans in the air with their fingers. You will be provided with a video of a person drawing in the air, carefully analyze the video to determine what letter or digit the person draws. Only respond with the character detected, no explanation. The only valid responses are a letter or a digit. `ASSISTANT`:

A: A

VideoLLaMA2

Q: You are a video classifier trained to detect letters and digits drawn by humans in the air with their fingers. You will be provided with a video of a person drawing in the air, carefully analyze the video to determine what letter or digit the person draws. Only respond with the character detected, no explanation. The only valid responses are "doing other things" or "doing nothing" or a letter or a digit.

A: doing nothing

VideoLLaMA2 (w/o contrast classes)

Q: You are a video classifier trained to detect letters and digits drawn by humans in the air with their fingers. You will be provided with a video of a person drawing in the air, carefully analyze the video to determine what letter or digit the person draws. Only respond with the character detected, no explanation. The only valid responses are "doing other things" or "doing nothing" or a letter or a digit.

A: 2

Fig. 8: Vision Language Model Evaluation. We show the prompts we use to evaluate vision-language models.

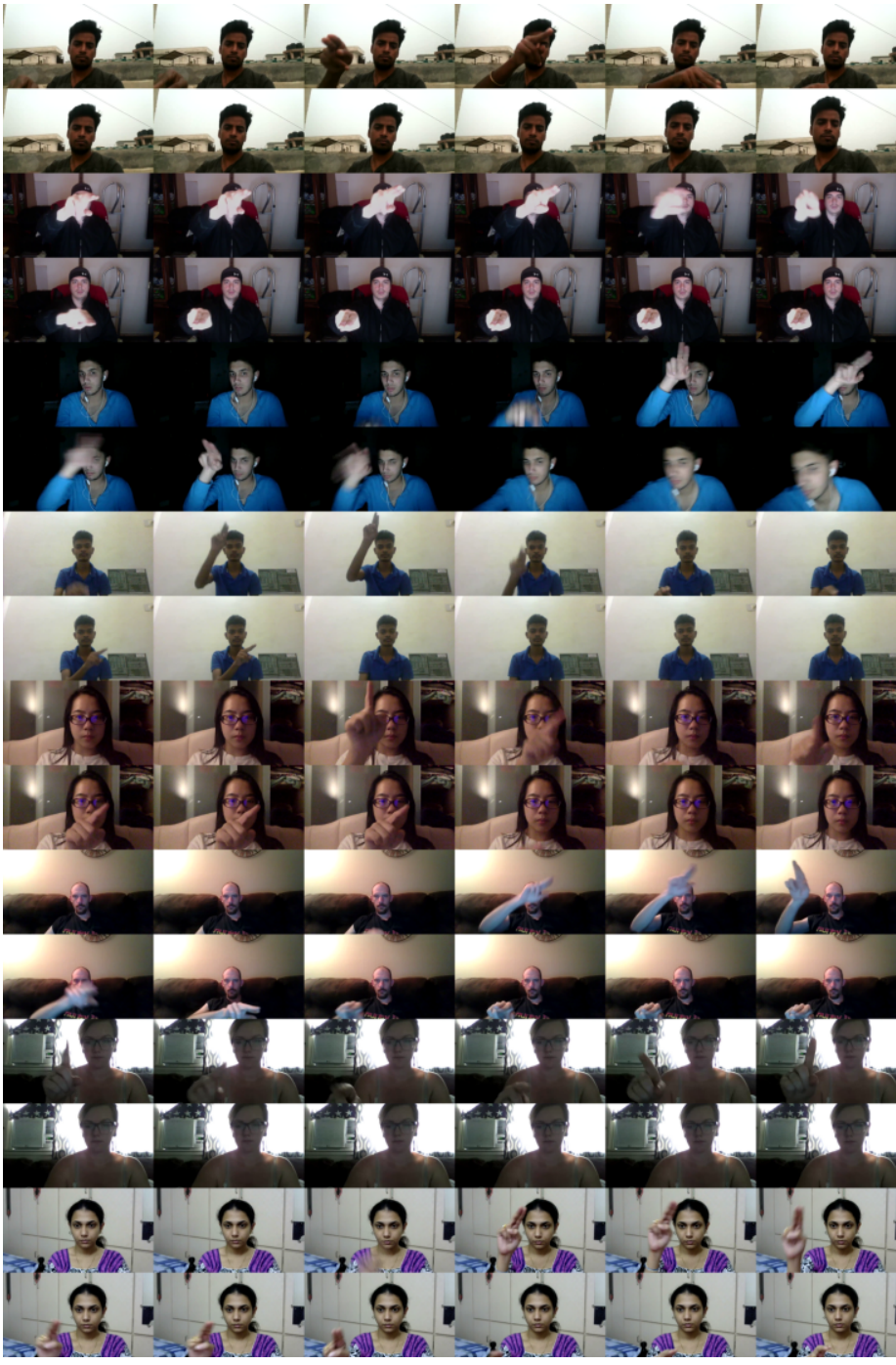

Fig. 9: Additional Examples. We show a few more video examples from our dataset by sampling 12 frames uniformly from randomly selected videos.

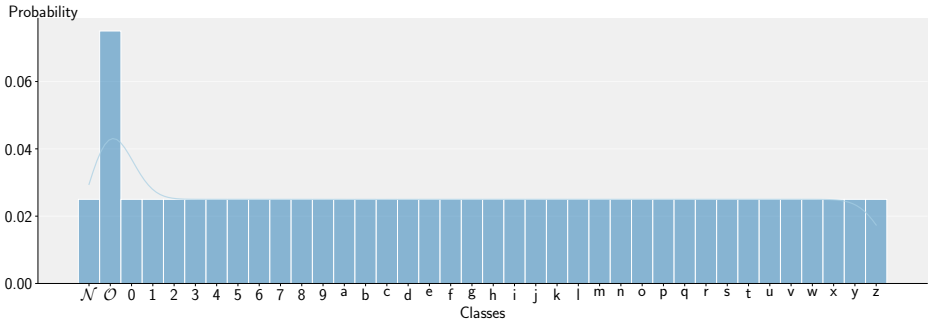

Fig. 10: Distribution over classes in the AirLetters dataset.

C Datasheet

We present a datasheet for our dataset inspired by the template in [32].

C.1 Motivation

For what purpose was the dataset created?
The purpose of the AirLetters dataset is to support training and evaluation of the ability of models to recognize motion patterns and to perform temporal aggregation of information across a video.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by the authors of the paper on behalf of Qualcomm Technologies Inc. and TwentyBN GmbH.

Who funded the creation of the dataset?
The creation of this dataset was funded by Qualcomm Technologies Inc. and TwentyBN GmbH.

Any other comments?
No.

C.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
An instance of the dataset consists of a video of a person drawing a character in the air with their hands, as well as a corresponding label.

How many instances are there in total (of each type, if appropriate)?

There are 161652 videos in total. Each of the classes has an equal number of videos except for the class "doing other things" which has 3× the number of videos.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
This dataset contains all possible instances and is not a sample from a larger set.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
An instance of the dataset consists of a video as well as the following information:

Worker ID. A unique integer worker ID to differentiate between individuals who recorded the videos.
Duration. Duration of the video.
Label. The letter or digit drawn by the worker using their hand(s) or one of the contrast classes: doing nothing or doing other things.

Is there a label or target associated with each instance? If so, please provide a description.
A label can be a letter, a digit, or one of the contrast classes: "doing nothing" or "doing other things".

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
N/A.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
The dataset has an 8:1:1 split into training set, validation set and test set.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Videos were reviewed to detect potential errors, but it is not guaranteed from being free of any errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, the dataset contains videos of humans drawing characters in the air with their hands.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
While the faces of the individuals in the video are visible, the videos were collected under a direct agreement with the crowd workers, permitting research and commercial use. The audio and meta-data information from the videos was removed.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, this dataset does not contain sensitive data.

Any other comments?
No.

C.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The dataset was collected with the help of crowdworkers and contractors.

A simple web interface was used for recording videos and creating annotations. The resulting data was manually inspected to ensure data integrity.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
N/A

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
N/A.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, the dataset contains videos of humans drawing characters in the air with their hands.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was collected directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Yes.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, crowdworkers signed a consent form.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Yes, the participants may reach out to us via email:
`research.datasets@qti.qualcomm.com`.

C.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Based on the aspect ratio of the originally recorded video, we perform an aspect-ratio preserving resizing to the width off 640 pixels. As a result, all videos are either (360, 640) or (480, 640). All of videos are preprocessed to have a framerate of 30 FPS.

Was the "raw" data saved in addition to the preprocessed/cleaned /labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No.

C.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Yes. In this work, baseline models are evaluated on the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Currently, such a repository does not exist.

What (other) tasks could the dataset be used for?
The dataset can be also be used for pre-training video understanding or video generation models.

C.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, we plan to make the dataset publicly available.

How will the dataset be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset will be publicly downloadable through a website.

When will the dataset be distributed?
The dataset should be available publicly on the dataset website.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
Yes, we plan to release the dataset under a proprietary research license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

C.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset is hosted and maintained by Qualcomm Technologies Inc.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The owners of the dataset can be contacted through:
`research.datasets@qti.qualcomm.com`.

Is there an erratum? If so, please provide a link or other access point.
N/A

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

If the dataset is updated, changes should be communicated through the dataset web page.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
N/A.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Certain mechanisms exists for research use cases. Further information is detailed in the proprietary research license.