

EHWGesture - A dataset for multimodal understanding of clinical gestures

Gianluca Amprimo^{1*} Alberto Ancilotto² Alessandro Savino¹ Fabio Quazzolo¹
Claudia Ferraris³ Gabriella Olmo¹ Elisabetta Farella² Stefano Di Carlo¹
¹Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy
²Fondazione Bruno Kessler, Trento, Italy
³CNR-IEIT, Torino, Italy

name.surname@polito.it, {aancilotto, efarella}@fbk.eu, claudia.ferraris@cnr.it

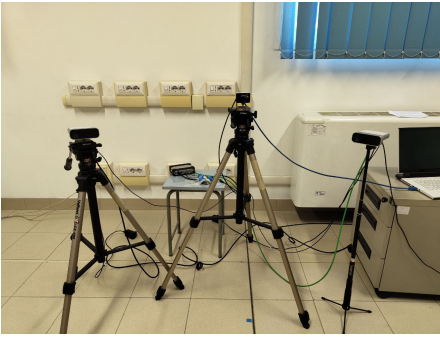


Figure 1. Acquisition setup. Orthogonal Azure Kinect cameras and the DVXplorer Lite event camera in the middle.

1. Dataset evaluation

The dataset is approximately 250 GB in size. As supplementary material, we provide the trials recorded by five subjects (around 50 GB), while the remaining 20 subjects' data will be fully released upon the manuscript's final acceptance.

2. Data collection Setup

Figure 1 illustrates the setup used for data collection.

3. Data preprocessing

To limit computational complexity when training classification models for both tasks, we applied temporal and spatial downsampling to input videos. For gesture and quality classification, the network processes frame windows of varying sizes using a variable frame rate, as analyzed in the main paper. The spatial resolution of inputs is reduced through cropping and resizing operations. Specifically, we use `mediapipe-hands` to detect hand bounding boxes within each frame, removing irrelevant regions that do not

Table 1. Effect of different pretraining approaches.

Task	No pretraining	NTA	PTA
AQA	76.16	79.12	78.93
Gesture	89.06	93.72	93.92
Combined	67.84	74.15	74.12

contribute to gesture execution. Figure 2 illustrates this cropping process for the `Main` and `Sub` camera videos. The resulting cropped frames are then resized, maintaining a resolution of 240 pixels on the longer side. Event data is aggregated into frames using synchronization triggers received by the DVXplorer Lite device at the start of each exposure cycle of the Kinect cameras. Aggregation involves accumulating events occurring at the same pixel locations, separately for each polarity, forming a bi-dimensional matrix. Before accumulation, noise reduction is performed by filtering out isolated events within a 10 ms temporal window. Additional noise is removed by setting pixels with simultaneous positive and negative polarity to zero. As with RGB data, a heatmap is generated by analyzing the number of events per pixel across all recorded trials. The most active region, sized 150×150px, is then cropped and retained.

4. Ablation on network pretraining

We performed an ablation to test different pairing strategies in contrastive learning. Alongside our original setup (treating temporally misaligned samples as negatives - NTA), we also tested using temporally adjacent samples as positives (PTA). As shown in Tab 1, PTA slightly improves gesture recognition but slightly lowers AQA performance, suggesting choice may depend on the downstream task.

*Corresponding author.

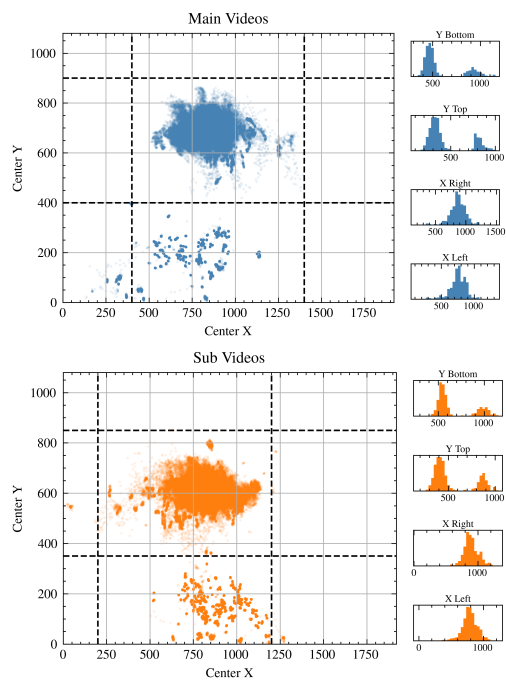


Figure 2. Detection locations across all videos for the Main and Sub cameras. The cropped region for training the gesture and quality classification networks is highlighted. Detections in the lower part of the frame refer to the hand not being used for gestures.