Supplementary Material:
LD-ConGR: A Large RGB-D Video Dataset for Long-Distance Continuous Gesture Recognition

Dan Liu, Libo Zhang, Yanjun Wu
1Institute of Software Chinese Academy of Sciences, Beijing, China
2Hangzhou Institute for Advanced Study, UCAS, Hangzhou, China
{liudan, libo, yanjun}@iscas.ac.cn

1. Recording Spots

Six recording spots are set in each scene, as shown in Fig. 1. The camera is fixed in front of the conference table. The 6 recording spots are evenly distributed on the left and right sides of the conference table, named L1, L2, L3, R1, R2 and R3 respectively from near to far from the camera. The distance between the recording spot and the camera is between 1m and 4m.

2. Gesture Region Estimation

In long-distance gesture recognition, the area where the gesture occurs is small compared to the background area. To reduce redundant information and focus on the gesture, we first estimate the gesture region based on the hand location and then conduct recognition in the estimated region. In the training stage, the hand location can be obtained from the annotations. During the test phase, a lightweight hand detector is used to locate the hand. Benefiting from the hand location annotations of LD-ConGR, the hand detector can be well-trained on LD-ConGR. Algorithm 1 and Algorithm 2 illustrate the specific processes of training and predicting with gesture region estimation. In our experiments, the YOLO V4 tiny [1] is adopted as the hand detector.

Algorithm 1 Gesture recognition model training with gesture region estimation.

**Input:** Dataset $D_{train} = \{S_1, S_2, ..., S_n\}$. Each gesture sample $S_i \in D_{train}, i \in \{1, 2, ..., n\}$ is given the frame sequence $F_i = [f_{i,1}, f_{i,2}, ..., f_{i,m}]$, the gesturing hand location $(x_{i,j}, y_{i,j}, w_{i,j}, h_{i,j})$ in each frame $f_{i,j}, j \in \{1, 2, ..., m\}$, and the category $c_i$

**Output:** Well-trained gesture recognition model weights $M_g$; hand detector $M_h$

1. $D'_{train} = \emptyset$
2. for all $S_i \in D_{train}$ do
3. Get the hand location $R_{hand} = (x_{i,1}, y_{i,1}, w_{i,1}, h_{i,1})$ of the first gesture frame $f_{i,1}$ from the annotations
4. $R_{ges} = (x_{i,1}, y_{i,1}, 5 \times w_{i,1}, 4 \times h_{i,1})$ ⊳ Estimated gesture region
5. $F'_i = []$
6. for all $f_{i,j} \in F_i$ do
7. $f'_{i,j} = Crop(f_{i,j}, R_{ges})$
8. Append $f'_{i,j}$ to $F'_i$
9. end for
10. Add $S'_i = (F'_i, c_i)$ to $D'_{train}$
11. end for
12. Train model $M_g$ on $D'_{train}$
13. Train a tiny hand detector $M_h$ based on hand instances in $D_{train}$
14. return $M_g, M_h$
Algorithm 2 Continuous gesture prediction with gesture region estimation.

**Input:** Test video $v$; Gesture recognition model $M_g$; Hand detector $M_h$

**Output:** Gesture predictions

1: $C_{\text{hand}} = \{(\text{id}, \text{region list}, r_{bbox}, f_{id})\}$ \text{ hand cache}
2: $\text{cur}_\text{frame} = \text{Read}(v)$
3: \textbf{while} $\text{cur}_\text{frame}$ do
4: \hspace{1em} Detect hands on $\text{cur}_\text{frame}$ with hand detector $M_h$, $\text{detections} = \{(\text{id}, h_{bbox})\}$
5: \hspace{1em} \textbf{for all} $h_i \in \text{detections}$ do
6: \hspace{2em} if $h_i$ matches $h_j \in C_{\text{hand}}$ \text{ based on hand locations} then
7: \hspace{3em} Estimate gesture region $r_t$ based on $h_{bbox}$ of $h_i$
8: \hspace{3em} Update latest matched frame $f_j$ to $\text{cur}_\text{frame}$ for hand $h_j$
9: \hspace{3em} Update region $r_{bbox}$ to $r_t$ for hand $h_j$
10: \hspace{3em} Crop $r_t$ from $\text{cur}_\text{frame}$ and add it to the $\text{region list}$ of hand $h_j$
11: \hspace{1em} \hspace{1em} \textbf{else}
12: \hspace{2em} Add the new hand instance $h_i$ to $C_{\text{hand}}$
13: \hspace{1em} \hspace{1em} \textbf{end if}
14: \hspace{1em} \textbf{end for}
15: \hspace{1em} Do gesture recognition on all the $\text{region list} \in C_{\text{hand}}$ with the well-trained model $M_g$
16: \hspace{1em} $\text{cur}_\text{frame} = \text{Read}(v)$
17: \textbf{end while}
18: \textbf{return} Gesture prediction results

3. Ethics Statement

The data is only allowed for academic research and we will provide strict access for applicants who sign data use agreements. The subjects involved in data collection were informed of the uses of the data and signed informed consent.

References