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

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

1. Recording Spots

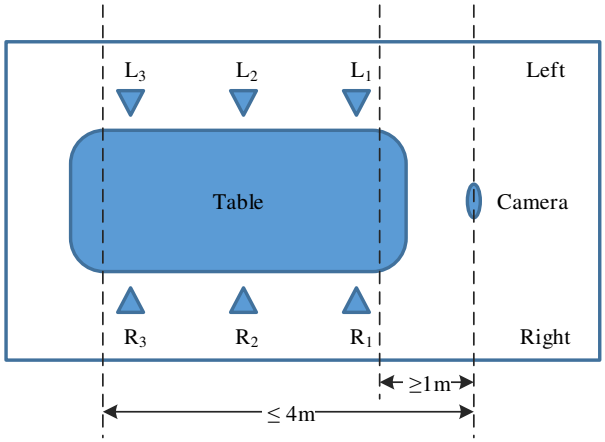

Figure 1. Six recording spots are set in each scene. The distance between the recording spot and the camera is between $1m$ and $4m$.

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 L_1, L_2, L_3, R_1, R_2 and R_3 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

$D'_{train} = \emptyset$
for all $S_i \in D_{train}$ **do**
  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 ▷ $(x_{i,1}, y_{i,1})$ is the center coordinates of the hand bounding box.
  $R_{ges} = (x_{i,1}, y_{i,1}, 5 \times w_{i,1}, 4 \times h_{i,1})$ ▷ Estimated gesture region
  $F'_i = []$
  for all $f_{i,j} \in F_i$ **do**
    $f'_{i,j} = Crop(f_{i,j}, R_{ges})$
    Append $f'_{i,j}$ to F'_i
  end for
  Add $S'_i = (F'_i, c_i)$ to D'_{train}
end for
Train model M_g on D'_{train}
Train a tiny hand detector M_h based on hand instances in D_{train}
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

$C_{hand} = \{(h_{id}, region\_list, r_{bbox}, f_{id})\}$ ▷ hand cache
$cur\_frame = Read(v)$
while cur_frame **do**
  Detect hands on cur_frame with hand detector M_h, $detections = \{(h_{id}, h_{bbox})\}$
  for all $h_i \in detections$ **do**
    if h_i matches $h_j \in C_{hand}$ ▷ based on hand locations **then**
      Estimate gesture region r_t based on h_{bbox} of h_i
      Update latest matched frame f_j to cur_frame for hand h_j
      Update region r_{bbox} to r_t for hand h_j
      Crop r_t from cur_frame and add it to the $region\_list$ of hand h_j
    else
      Add the new hand instance h_i to C_{hand}
    end if
  end for
  Do gesture recognition on all the $region\_list \in C_{hand}$ with the well-trained model M_g
  $cur\_frame = Read(v)$
end while
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

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1