

Self-Mutual Distillation Learning for Continuous Sign Language Recognition

Aiming Hao^{1,2}, Yuecong Min^{1,2}, Xilin Chen^{1,2}

¹Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China

²University of Chinese Academy of Sciences, Beijing, 100049, China

{aiming.hao,yuecong.min}@vip1.ict.ac.cn, xlchen@ict.ac.cn

A. Overview

In this supplementary material, we provide details and extra experiment that are not shown in the main paper. Firstly, we provide more visualization results about the similarity matrix (Sect. B). Then we introduce the details of the proposed GSBA method (Sect. C). At last, we give the extra experiment to explore the influence of the receptive field (Sect. D).

B. Visualization of the Similarity Matrix

In Sect.3.4, we give an instance to explore the property of the features after sharing the weights. As shown in Fig. 3, we also visualize the similarity matrices in different cases: without WS, with WS, and with WS & GSBA. We can find that only with the WS method, the GCF of a key frame will have a strong correlation with the LVFs of the frames nearby it. Meanwhile, with the gloss segmentation, the GCF of a key frame will focus on more LVFs nearby it. More visualization results on the training set and the development set are given in Fig. 4. And we find that all of them show a similar characteristic with the instance we gave in Sect.3.4.

C. Gloss Segment Boundary Assignment

The pseudo-code of GSBA is described in Algorithm 1. It consists of two functions: LOCATE and EXPAND.

For the LOCATE function, it is used to locate the key frame of the current class c_i . We scan the predict probability distribution \hat{Y}_g to get the location t .

For the EXPAND function, it is used to expand the key frame. We treat the located key frame as an anchor frame. We first set an expanded radius d to limit the maximum expansion distance and a direction $s \in \{-1, 1\}$ to determine the expand direction. Then, we expand the frame from $t + s$ to $t + s * d$ frame. If the cosine similarity between the GCF of the current expanding frame and the weight vector of c_i is the smallest among the classes $c_j \in L$, we then annotate this frame with the label c_i . Otherwise, we will stop the expansion process.

Algorithm 1 Gloss Segment Boundary Assignment

Input: video's GCF sequence G , ground truth sign gloss sequence l , predict probability distribution \hat{Y}_g , classifier weight vectors W , expanded radius d .

Output: pseudo gloss segment labels y^{seg}

```
1: function LOCATE( $l_i, pos$ )
2:   for  $t \leftarrow pos; t \leq |G|$  do
3:      $c_t \leftarrow \arg \min \{\hat{y}_t\}$ 
4:     if  $c_t == l_i$  then
5:        $pos \leftarrow t$ 
6:     return  $t, pos$ 
7:   else
8:     break
9:   end if
10: end for
11: end function
12:
13: function EXPAND( $t, l_i, s$ )
14:   for  $j \leftarrow 1; j \leq d$  do
15:      $c_{t+js} \leftarrow \arg \min \{\langle g_t, w_c \rangle\}_{c \in L}$ 
16:     if  $c_{t+js} == l_i$  then
17:        $y^{seg} \leftarrow \{t + js, l_i\}$ 
18:     else
19:       break
20:     end if
21:   end for
22: end function
23:
24: for  $i \leftarrow 1; i \leq |G|$  do
25:    $y^{seg} \leftarrow \{i, \text{blank}\}$ 
26: end for
27:  $pos = 1$ 
28: for  $i \leftarrow 1; i \leq |L|$  do
29:    $t, pos \leftarrow \text{LOCATE}(l_i, pos)$ 
30:    $y^{seg} \leftarrow \{t, l_i\}$ 
31:   EXPAND( $t, l_i, -1$ )
32:   EXPAND( $t, l_i, 1$ )
33: end for
34: return  $y^{seg}$ 
```

Based on the two functions mentioned above, we first initialize each frame’s label as the blank class. Then we use the LOCATE function to locate the key frames for each class l_i in the ground truth sign gloss sequence \mathbf{l} . After that, we use the EXPAND function to expand the key frames and update the pseudo gloss segment labels \mathbf{Y}^{seg} . Then we smooth the \mathbf{Y}^{seg} , and get the smoothed labels $\hat{\mathbf{Y}}^{seg}$ as:

$$\tilde{y}_{ij}^{seg} = \begin{cases} 1 - \varepsilon & \text{if } j = y_i^{seg} \\ \frac{\varepsilon}{|\mathbb{G}|+1} & \text{otherwise,} \end{cases} \quad (1)$$

where ε is the label smoothing rate. More visualizations of the pseudo gloss segment label produced by GSBA are shown in Fig. 5.

Note that, we active the GSBA after epoch 20 to avoid the unreliable segment proposal at the initial optimization stage. And we enlarge the expanded radius d after the training of the contextual module tends to steady to introduce more spatial-temporal information.

D. Details on Temporal Receptive Field

We define r as the temporal receptive field (TRF) of the visual module. As r is relevant to a temporal window in the contextual module, and suitable size of the visual module’s TRF will better match the followed contextual module. As shown in Fig. 1 and Fig. 2, we visualize the self-similarity matrices of the visual module with different TRFs and find that the self-similarity matrix with larger TRF tends to be more diagonal. The blank and non-blank features will be hard to distinguish if the TRF is small, and this will increase the difficulty of aligning features from the two modules. Moreover, we compare the performances with different TRF of the visual module as shown in Table 1. We observe that small r will results in performance deterioration due to the loss of temporal information. Besides, the overuse of the pooling operation will also harm the performance. Among the selected TRFs, the optimal structure is $C_5 - P_2 - C_5$. So we set the visual module as $C_5 - P_2 - C_5$ by default.

Table 1. Ablation studies on the TRF of the visual module on the PHOENIX14 dataset (only trained in the synchronous training stage).

Visual Module	TRF	Dev (%)	Test (%)
C_5	5	22.5	23.3
$C_5 - C_5$	9	22.4	23.2
$C_5 - P_2 - C_5$	14	21.2	21.4
$C_5 - P_2 - C_5 - P_2$	16	22.1	22.7

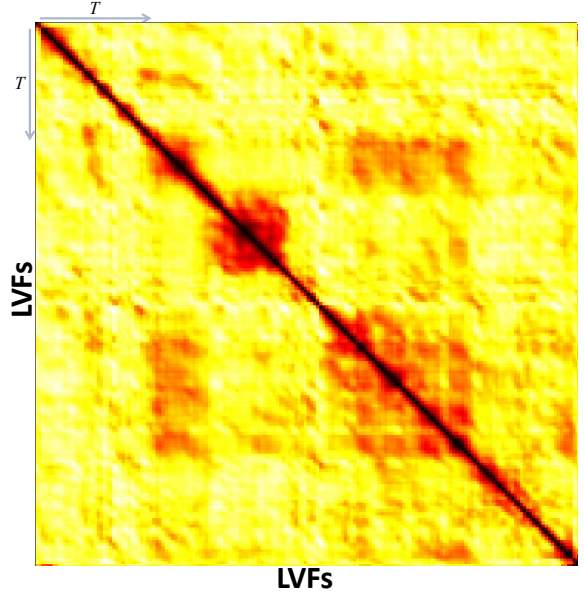


Figure 1. The heatmaps of LVFs’ self similarity matrix with receptive fields 5.

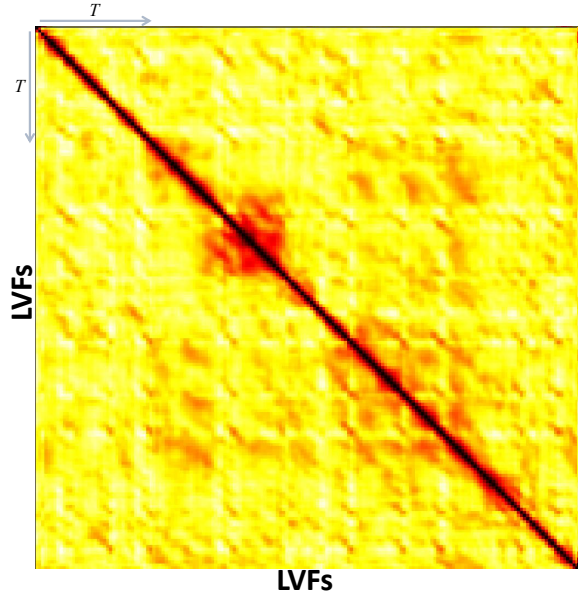
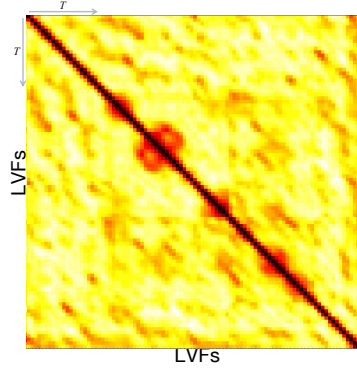
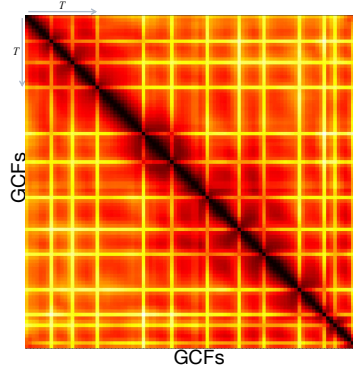


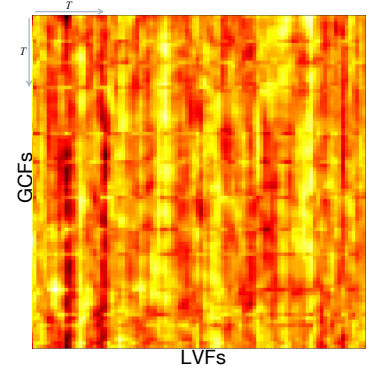
Figure 2. The heatmaps of LVFs’ self similarity matrix with receptive fields 9.



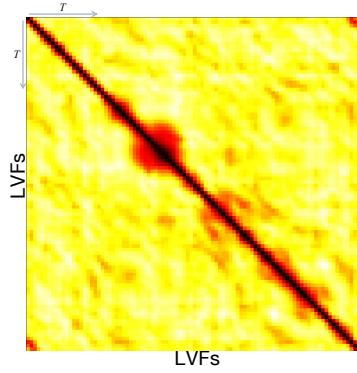
(a) Self similarity matrix of the LVFs.



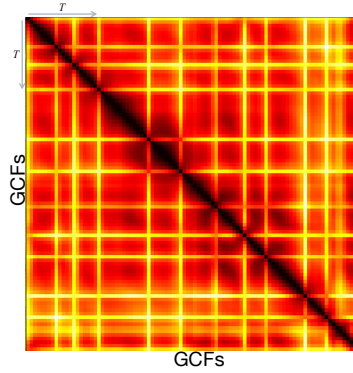
(b) Self similarity matrix of the GCFs.



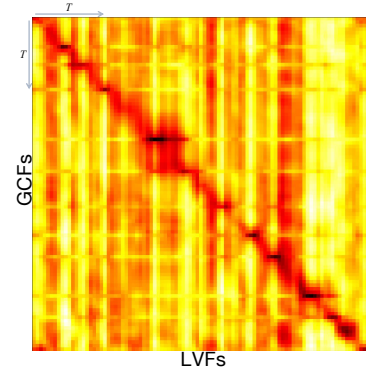
(c) Similarity matrix between the LVFs and GCFs.



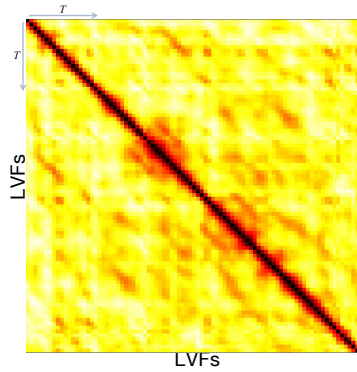
(d) Self similarity matrix of the LVFs.



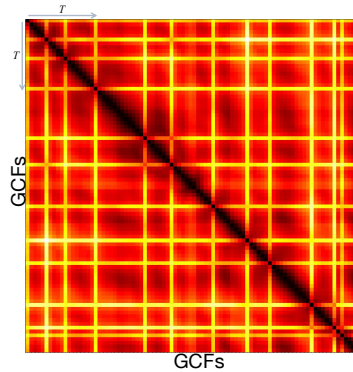
(e) Self similarity matrix of the GCFs.



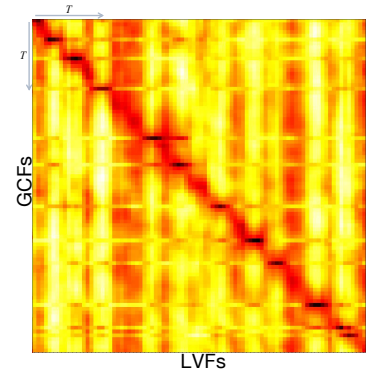
(f) Similarity matrix between the LVFs and GCFs.



(g) Self similarity matrix of the LVFs.

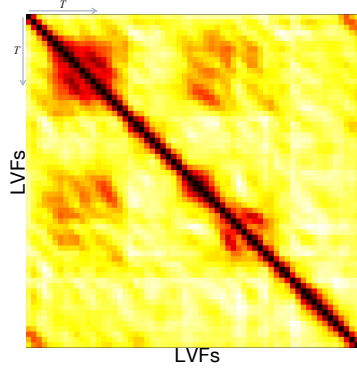


(h) Self similarity matrix of the GCFs.

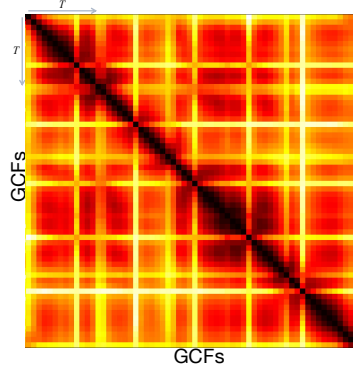


(i) Similarity matrix between the LVFs and GCFs.

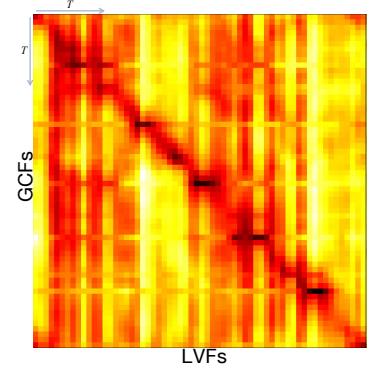
Figure 3. The heatmap of the LVFs' and GCFs' self-similarity matrices and the similarity matrix between the LVFs and the GCFs (**the darker color represents the higher similarity**). From top to bottom are the results that network training without WS, with WS and with WS & GSBA.



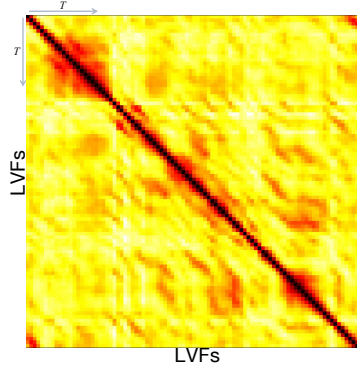
(a) Self similarity matrix of the LVFs.



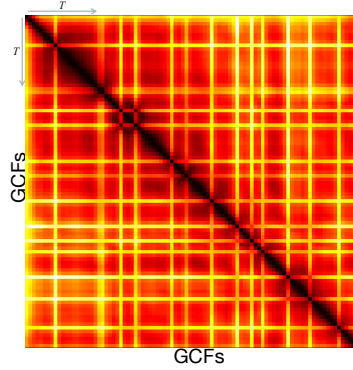
(b) Self similarity matrix of the GCFs.



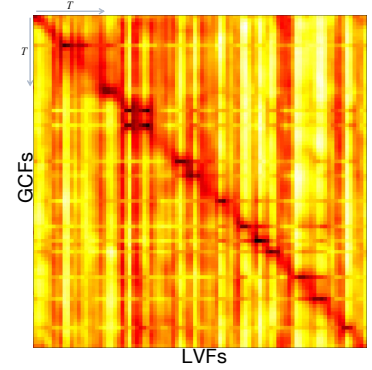
(c) Similarity matrix between the LVFs and GCFs.



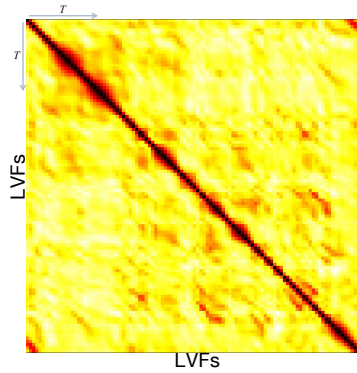
(d) Self similarity matrix of the LVFs.



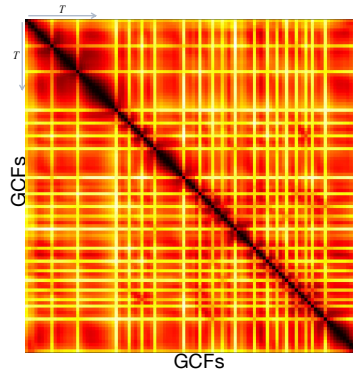
(e) Self similarity matrix of the GCFs.



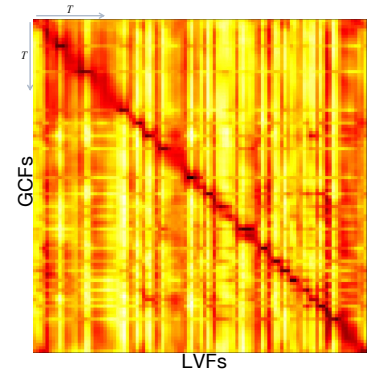
(f) Similarity matrix between the LVFs and GCFs.



(g) Self similarity matrix of the LVFs.



(h) Self similarity matrix of the GCFs.



(i) Similarity matrix between the LVFs and GCFs.

Figure 4. The heatmap of the LVFs' and GCFs' self-similarity matrices and the similarity matrix between the LVFs and the GCFs in different examples (the darker color represents the higher similarity).

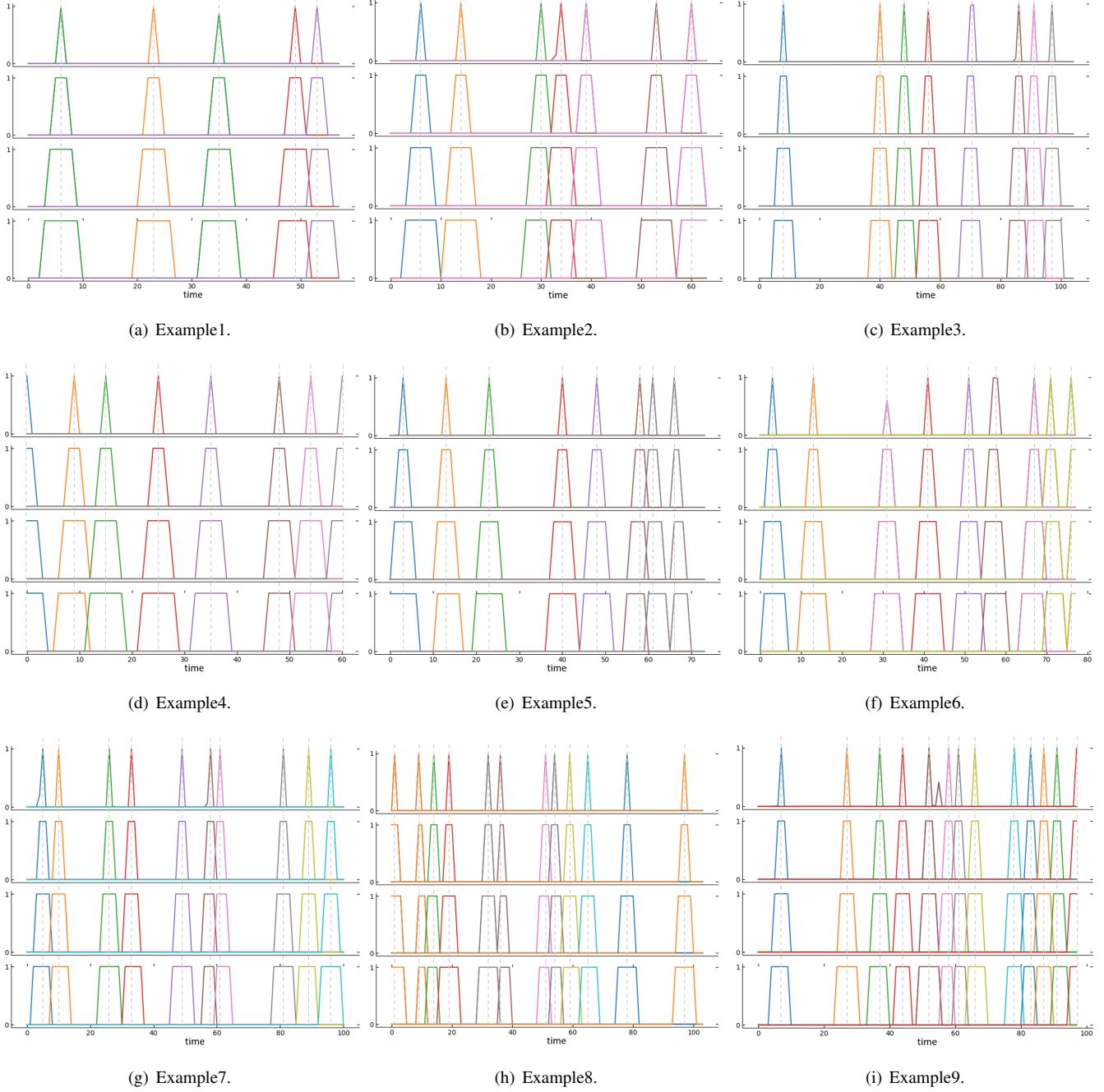


Figure 5. From top to bottom are the spike phenomenon and the pseudo gloss segment labels produced by GSBA with $d = 1, 2, 3$ (different colors represent different classes).