Supplementary Material for Score-level Multi Cue Fusion for Sign Language Recognition

Anonymous ECCV submission

Paper ID 16

1 Additional Information for the Section 4

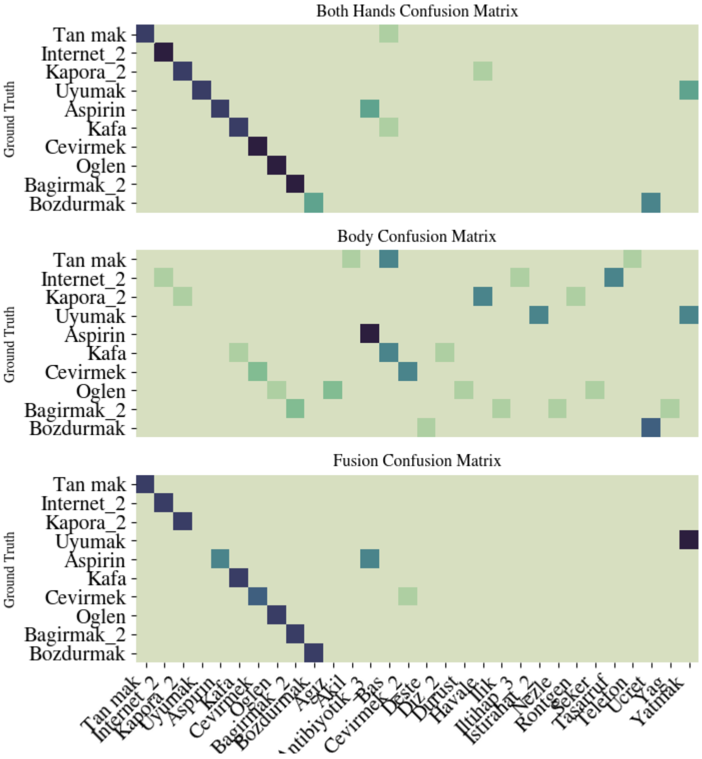

Fig. 1. Confusion matrix of the top ten classes where Hand model have advantage over the Body Model. Hand model performs successful for the given sign glosses, whereas body model fails to capture the difference. Multi-cue fusion model manages the capture most dense features from the bost models and successful except the SLEEP(v) gloss

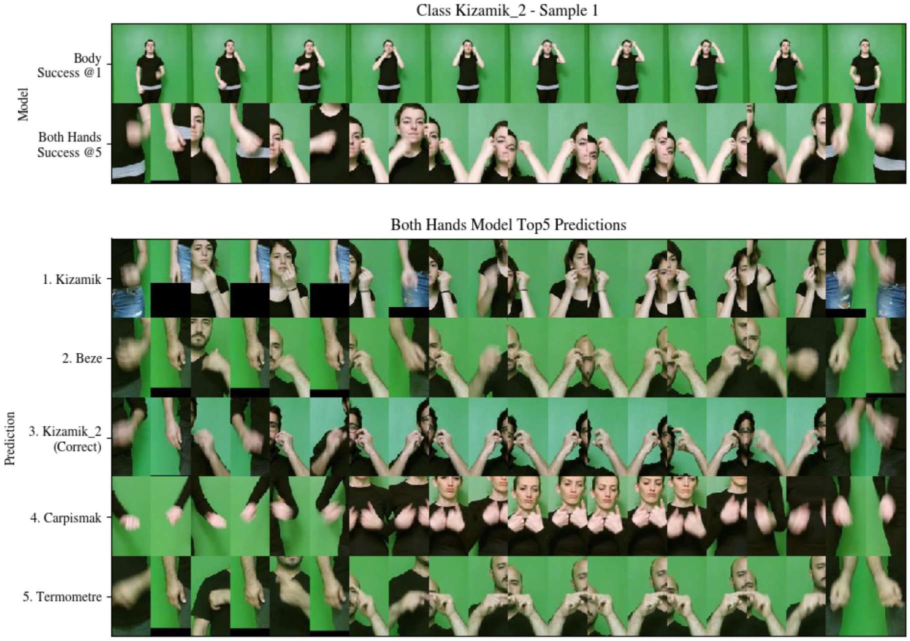

Fig. 2. Investigating hand cue failure settings on a MEASLES(n) gloss. Failed top predictions are performed for visualization.

Table 1. Top-N Model Accuracy for the Body, Hand and Face settings. Cue models have increasing accuracy values with the increasing N. Hand cue model perform best in all settings.

Setting	Top-1	Top-2	Top-3	Top-5	Top-10
Body	81.83	90.45	93.59	96.09	98.32
Hand	88.70	95.00	96.79	97.75	**98.76**
Face	37.00	46.49	51.35	57.91	65.92

2 Experiment details omitted in Section 3

2.1 Pose Keypoint Details

Thumb keypoint refers to the 9th point in Hand21 setting in [**?**].

2.2 Failed Settings

For the single hand setting, we have also tried to use the non-dominant hand, and for the mixed setting we have also experimented with adding a black padding alignment to single parts. Both attempts resulted in severe accuracy loss.

We have also experimented extending the best performing hand setting with the face crop. Face crops are horizontally concatenated to each hand setting. Horizontal crop size defined as $350/N$ pixels, where N is the number of the sub units in the image. Vertical crop size is remained as 350 pixels. Final image is resized to 114×114 as in the previous setting. This setting has resulted in 76.99% percent accuracy on the test set, with a 10.73% related performance drop compared to the hand counterpart. We suspect that the performance drop occurs due to decreasing spatial size with each cue included into single image, which causes a race condition between the each cue.

3 Complementary figures for the Spatio-Temporal Sampling

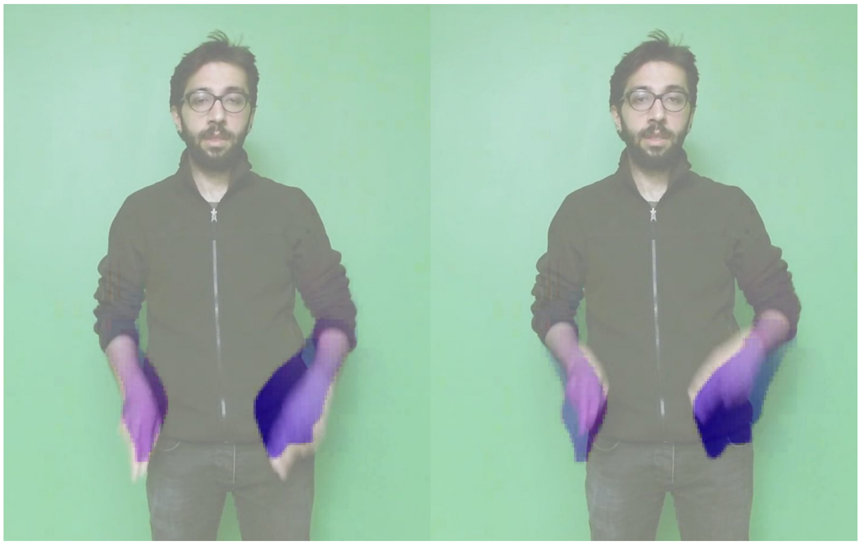

Fig. 3. A temporal cue detection visualization. Our prosed euclidian based hand tracking algorithm detects the visual activity and flow. Defined threshold value determines the strenght of the visual flow. Image is generated using the dense Optical Flow algorithm

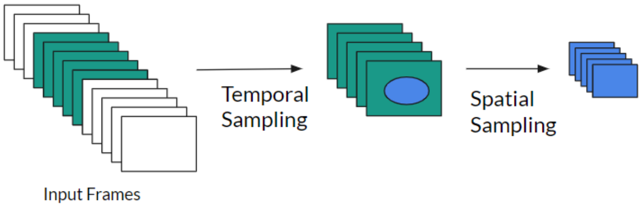

Fig. 4. Spatio-temporal sampling pipeline is visualized. Temporal sampling is followed by spatial sampling to generate trainable cue set