A data augmentation approach for sign-language-to-text translation in-the-wild

Fabrizio Nunnari - DFKI, Saarland Informatics Campus D3 2, Saarbrücken, Germany
Cristina España-Bonet - DFKI, Saarland Informatics Campus D3 2, Saarbrücken, Germany
Eleftherios Avramidis - DFKI, Alt Moabit 91c, 10559 Berlin, Germany

**Problem:** most SL-video-to-text translation systems focus on frontal view recognition of sign language performances in very controlled environments.

**However:** real settings are subject to different conditions: illumination, angle of view, cameras, clothing, background, skin tones, body proportions, ...

**Our Approach:** let specialized tools extract animation information and augment on a normalized and controlled environment.

---

**Option 1:** end-to-end translation (from pixels to text)

**Option 2:** (motion) feature extraction first
(pixels to features/animation, features/animation to text)

**Option 3 (ours):** motion data extraction and augmentation
(pixels to animation, augmentation, animation data to text)

**Advantages:**
- Leave specialized tools (e.g., MediaPipe, OpenPose), trained with much data, normalized environmental conditions (brightness, clothing, ...)
- Full control of the augmentation parameters (cameras, body sizes, execution speed, ...)

**Test** the model on Sign Language corpora recorded in un-controlled conditions.

**Hypothesis:** Despite training corpora are recorded in controlled environments, SL recognition will work better in non-controlled environment.

---

The research reported in this paper was supported by the BMBF (German Federal Ministry of Education and Research) in the project SOCIALWEAR (Socially Interactive Smart Fashion, DFKI Kst 22132).