SIGN LANGUAGE RECOGNITION VIA DEFORMABLE 3D CONVOLUTIONS AND MODULATED GRAPH CONVOLUTIONAL NETWORKS

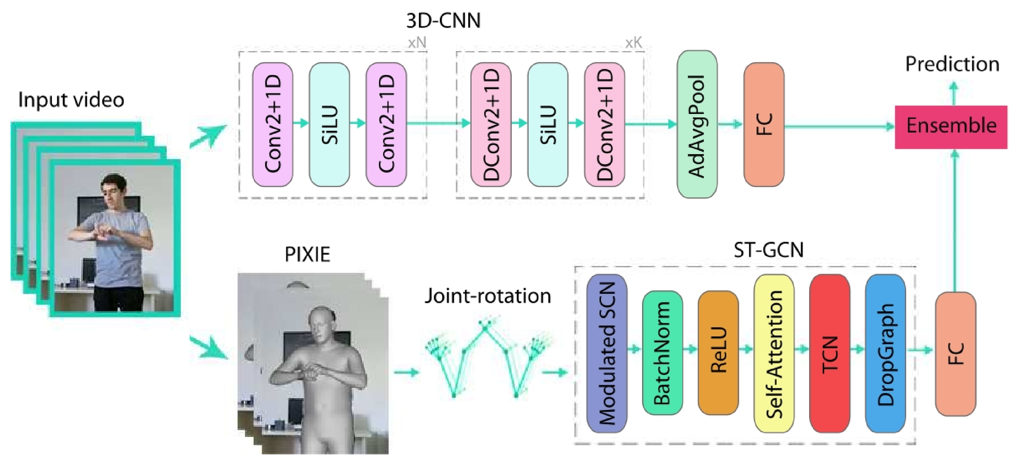
Katerina Papadimitriou   Gerasimos Potamianos

Department of Electrical & Computer Engineering, University of Thessaly, Volos, Greece

Overview

- **Goal**:
  - ✓ **Isolated sign language recognition (ISLR)** from videos in **signer-independent (SI)** mode.
- **Challenges:**
  - ✓ Strongly **correlated manual/non-manual** modalities.
  - ✓ **Inter-personal signing variation**.
- **Previous work [1]:**
  - ✓ Handshapes/mouthing **optical flow**, **skeletal**, and **appearance feature fusion**.
  - ✓ **Attentional encoder-decoder** with **temporal deformable convolutions** for **sign recognition**.
- **Paper contributions:**
  - ✓ **3D-CNN** model based on **deformable spatial** and **temporal convolutions**.
  - ✓ **Spatio-temporal graph convolutional network (ST-GCN)** relying on **modulated GCNs** [2].
  - ✓ **Graph construction** using **3D joint-rotation parameterization**.
- **Results:**
  - ✓ **Experiments** on a **Turkish** and a **Greek ISLR dataset**.
  - ✓ Achieve **new state-of-the-art** on Greek corpus and **competitive performance** on Turkish.

Overview of proposed ISLR system



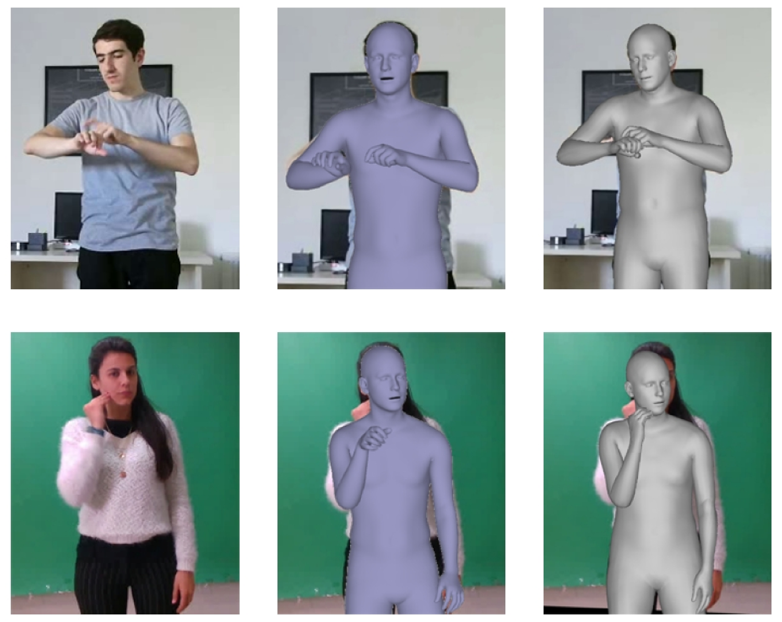
- **RGB frame modality:**
  - ✓ **3D-CNN** for **feature extraction** from **RGB** video frames.
    - o **Decouples spatial** and **temporal convolutions**.
    - o **Integrates deformable spatial** and **temporal convolutions**.
- **Skeleton sequence modality:**
  - ✓ **Graph construction:** "PIXIE" **3D joint-rotation parameterization** of the human **skeleton**.
  - ✓ **Attention-based ST-GCN: modulated GCNs** followed by **temporal convolutions**.
- **Ensemble module:**
  - ✓ **Fuse posteriors** from the last fully-connected layers of the **two** different **modalities**.

Our Approach (I)

- **Deformable 3D-CNN for RGB modality:**
  - ✓ **Backbone model:** 18-layer **ResNet2+1D** network [4].
    - o **3D convolutional kernels**   **spatial convolutional filters** and **temporal** ones.
  - ✓ **Replace spatial** and **temporal convolutions** with **deformable** counterparts.
  - ✓ **Deformable convolutions:**
    - o **Convolutional layer** to **predict** the **position offsets**.
    - o **Augment sampling grid** by **adding** the predicted **offsets** to **convolution**.
  - ✓ **Apply deformable** spatial and temporal **convolutions** in the **last 3** network **stages**.
  - ✓ **Replace** the **ReLU activation** function with the **SiLU** one.
- **Implementation details:**
  - ✓ **Crop** upper **body** using the **3D joints** generated by **MediaPipe** [5].
  - ✓ **Resize** to **256x256**.
  - ✓ **Pretrain** our model on the **Chinese SL dataset** [6].

Our Approach (II)

- **Modulated ST-GCN for skeletal modality:**
  - ✓ **ST-GCN unit** involves a **spatial GCN** followed by a **temporal convolution**.
  - ✓ **Employ modulated GCN**:
    - o **Weight modulation**: learnable weight modulation vector to modulate the weight matrix.
    - o **Affinity modulation**: adds a learnable mask to the adjacent matrix.
  - ✓ **Self-attention**: involves a spatial, a temporal, and a channel attention module.
  - ✓ **DropGraph** [7]: one node dropped together with its neighbor node set.
  - ✓ **10 modulated ST-GCN units** are utilized, followed by a **global average pooling layer**.
- **Graph construction:**
  - ✓ **3D joint-rotation parameterization** of the human pose as **graph feature representations**.
  - ✓ **"PIXIE":** infers **3D body pose** and **shape** parameterization using a moderator.
  - ✓ Regresses parameters for the **human shape** and **pose**, as well as the **facial expressions**.
  - ✓ **55 joints** with **6 degrees of freedom** ➡ 25 body pose joints and 15 joints per each hand.
  - ✓ **6x55**-dimensional **feature vectors**.
- **3D body reconstruction via "ExPose" regression model [8] (2nd column) vs "PIXIE" estimator (3rd column).**



- **Multi-modal Fusion:**
  - ✓ **Posteriors** from the **two** different **modalities** are appropriately **fused**.
  - ✓ **Assign** different **weights** to each **modality** in **accordance** with their individual **performance**.

Datasets & Experimental Setup

- **AUTSL dataset [9]:**
  - ✓ **226 Turkish isolated signs** performed by **43 signers**.
  - ✓ **36,302 RGB+D** videos in **20** different **backgrounds**.
  - ✓ **Official SI data split**:
    - o **28,142 training** videos **(31 signers)**.
    - o **4,418 validation** videos **(5 signers)**.
    - o **3,742 test** videos **(7 signers)**.
- **ITI GSL database [10]:**
  - ✓ 15 continuous **Greek SL** dialogues performed by **7** different **signers**, **5 times** each.
  - ✓ **40,826 isolated sign videos** with **vocabulary** size equal to **310**.
  - ✓ **RGB stream: 30 Hz rate, 648x480 resolution**.
  - ✓ **SI SLR** via **7-fold cross-validation**.
    - o **One test signer** per fold, with SLR models **trained** on the remaining **6**.

References

[1] **Papadimitriou & Potamianos,** "Multimodal sign language recognition via temporal deformable convolutional sequence learning," *Proc. Interspeech*, **2020**.
[2] **Zou & Tang,** "Modulated graph convolutional network for 3D human pose estimation," *Proc. ICCV*, **2021**.
[3] **Feng et al.,** "Collaborative regression of expressive bodies using moderation," *Proc. 3DV*, **2021**
[4] **Tran et al.,** "A closer look at spatiotemporal convolutions for action recognition," *Proc. CVPR*, **2018**.
[5] **Lugaresi et al.,** "MediaPipe: A framework for building perception pipelines," *Proc. CoRR*, **2019**.
[6] **Zhang et al.,** "Chinese sign language recognition with adaptive HMM," *Proc. ICME*, **2016**.
[7] **Cheng et al.,** "Decoupling GCN with DropGraph module for skeleton-based action recognition," *Proc. ECCV*, **2020**.
[8] **Choutas et al.**, "Monocular expressive body regression through body driven attention," *Proc. CVPR*, **2020**.
[9] **Camgöz et al.**, "Sign language transformers: Joint end-to-end sign language recognition and translation," *Proc. ICCV*, **2019**.
[10] **Huang et al.**, "Video-based sign language recognition without temporal segmentation," *Proc. AAAI*, **2018**.

Experimental results

- **Ablations on the introduced deformable 3D-CNN model:**
  - ✓ Ours achieves **95.39%** and **97.12% accuracies** on **AUTSL** and **ITI GSL.**

CNN Models	AUTSL	ITI GSL
C3D [27]	81.95	85.66
I3D [14]	87.64	89.11
P3D [28]	90.57	92.14
R3D [29]	92.04	94.03
ResNet2+1D + ReLU [17]	93.26	95.89
ResNet2+1D + SiLU	93.85	95.98
ResNet2+1D (pretrained) + SiLU [8]	94.77	96.51
Ours	**95.39**	**97.12**

- **Ablations on the proposed ST-GCN:**
  - ✓ **Important contributors** ➡ **Modulated GCN** and **attention mechanism**.
  - ✓ **"PIXIE" joint-rotation parameterization** ➡ **Highest recognition accuracies**.

ST-GCN Variations	AUTSL	ITI GSL
w/o Attention	93.88	94.85
w/o Modulated GCN	94.59	95.04
w/o DropGraph	95.12	95.79
w Decouple GCN	95.17	95.84
Ours	**95.32**	**96.14**

Streams	AUTSL	ITI GSL
2D Joint-position	94.96	95.46
3D Joint-position	95.10	95.68
2D Joint-motion	92.54	93.11
3D Joint-motion	93.24	93.57
Joint-rotation ("ExPose")	95.15	95.74
Joint-rotation ("PIXIE")	**95.32**	**96.14**

  - ✓ **3D-CNN appearance module outperforms the skeletal ST-GCN one.**
- **System evaluation** against literature ➡ both **modalities fusion** considered:
  - ✓ **ITI GSL :** 97.85% accuracy, **outperforming** the **state-of-the-art** (53% relative error reduction).
  - ✓ **AUTSL:** 96.67% accuracy, **trailing** the **state-of-the-art result** .
  - ✓ **Fusion improves** performance **over appearance** stream alone:
    - o 28% relative error reduction on AUTSL
    - o 25% relative error reduction on ITI GSL.

Dataset	Model	Modalities	Acc. (%)
AUTSL	VTN-PF [7]	A + HA + S	92.92
	MS-G3D [6]	A + S	96.15
	Ours	A + S	96.67
	SAM-SL [8]	A + S + F	**98.42**
ITI GSL	I3D + BiLSTM [14]	A	89.74
	OpenHands [31]	S	95.40
	Ours	A + S	**97.85**

Conclusions

- **Proposed** a **deep learning** model for **SI ISLR** from **RGB videos:**
  - ✓ **Integration** of **deformable convolutions** in the **ResNet2+1D network**.
  - ✓ **ST-GCN** ➡ **modulated GCNs**, **attention mechanism**, and **temporal convolutions**.
  - ✓ **Graph construction** using **3D joint-rotation parameterization** ➡ **"PIXIE"** approach.
  - ✓ **Fuse** both **modalities** in the **proposed system**.
- **Investigated** the **contribution** of:
  - ✓ **Fusing two** different **modalities** operating on **visual representations** of **appearance** and **human pose** to capture signing activity.
- **Achieved:**
  - ✓ **Competitive performance** on **AUTSL** dataset.
  - ✓ **New state-of-the-art** on the **ITI GSL** corpus.

Acknowledgments

The **research work** was supported by the **Hellenic Foundation for Research and Innovation (H.F.R.I.)** under the **"First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant"** (HFRI-FM17-2456).