4541

SPE-85.2

SPATIO-TEMPORAL GRAPH CONVOLUTIONAL NETWORKS FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

Maria Parelli [1] Katerina Papadimitriou [2] Gerasimos Potamianos [2] Georgios Pavlakos [3] Petros Maragos [1]

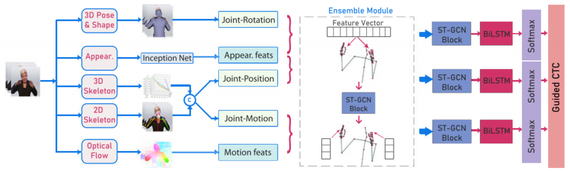
[1] School of Electrical & Computer Eng., National Technical University of Athens, Greece
[2] Dept. of Electrical & Computer Eng., University of Thessaly, Volos, Greece
[3] Electrical Eng. & Computer Sciences, University of California, Berkeley, CA, U.S.A.

icassp 2022 Singapore

Overview

- **Goal:**
  - ✓ **Continuous sign language recognition (CSLR)** from **RGB videos.**
- **Challenges:**
  - ✓ **Multitude**, **complexity**, and strong **correlation** of **SL articulators**.
  - ✓ **No** prior knowledge of **gloss level segmentation.**
- **Previous work [1]:**
  - ✓ Human **skeleton**, optical **flow**, and **appearance representations** fusion.
  - ✓ **Temporal deformable convolutional** based **sequence prediction**.
- **Paper contributions:**
  - ✓ **Multiple modalities:** signer's **pose**, **shape**, **appearance**, and **motion** information.
  - ✓ **Graph convolutional networks (GCNs)** with **bidirectional long short-term memory networks (BiLSTMs)**.
- **Results:**
  - ✓ **Experiments** on a **German** and a **Chinese CSLR dataset**.
  - ✓ Achieve **new state-of-the-art** on Chinese corpus and **competitive performance** on German.

Overview of proposed CSLR system



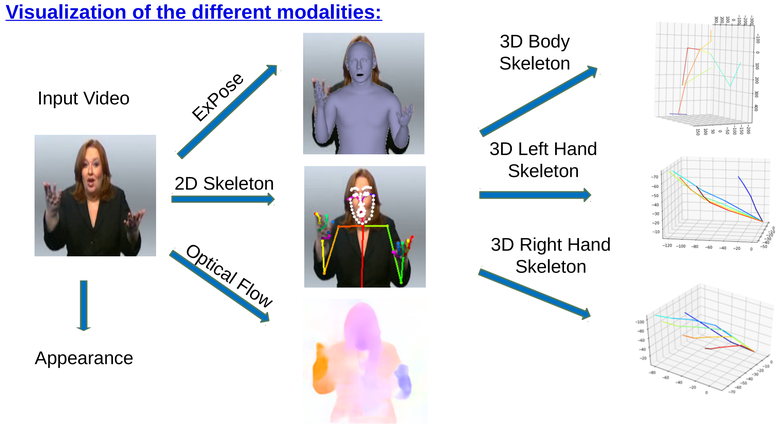
- **Visual features:**
  - ✓ **3D human pose** and **shape** extraction directly from RGB via "**ExPose**" [2].
  - ✓ **3D body/hand skeleton** regression through a **multi-layer neural network**.
  - ✓ **2D human pose** features generation via **OpenPose** [3].
  - ✓ **Appearance** and **optical flow** features.
- **ST-GCN ensemble:**
  - ✓ **Per-vertex feature vectors:** visual latent representations embedded with skeleton graphs.
- **Sequence learning model:**
  - ✓ **Spatio-temporal GCNs (ST-GCNs)** [4] with **BiLSTMs** for each feature vector.
  - ✓ Output **fusion** via a **guided CTC** approach for **gloss prediction**.

Visual Features (I)

- **3D human body pose and shape representation via "ExPose":**
  - ✓ **53 human-pose joints** with 6-dim **rotation representation**.
  - ✓ **10** face, body, and hands **shape** and **10 facial expressions** parameters.
- **2D human pose estimation via OpenPose:**
  - ✓ **137 2D human skeletal joints.**
  - ✓ Employ **15 body-pose keypoints** and **21 joints** per **hand**.
  - ✓ **Skeletal joints normalization**.
- **3D human body and hand skeleton regression:**
  - ✓ **"Lifting" 2D skeletal joint** locations to the **3D space** via a **regression model** [5].
  - ✓ **2** different **models** for 3D **upper-body** and **hands pose** estimation.
  - ✓ **17 3D upper-body pose** landmarks and **21 3D joints** coordinates **per hand**.
- **Appearance and optical flow features:**
  - ✓ **Appearance representation** via **Inception Net**, yielding **1024-dim** features.
  - ✓ **Motion informative image** generation via **SpyNet** ➡ **512-dim** features via **AlexNet**.

Visual Features (II)

- **Visualization of the different modalities:**



CSL Recognizer

- **Features blending via ST-GCN, producing 3 feature maps:**
  - ✓ **2D/3D skeletal-joint positions** embedded with **appearance features**.
  - ✓ **2D/3D joint-motion vector** embedded with **optical flow features**.
  - ✓ **"ExPose" representation** embedded with **appearance features**.
- **Graph and RNN based SL recognition:**
  - ✓ **ST-GCN unit**: **graph** and **temporal convolution**, capturing **short-term dynamics**.
  - ✓ **BiLSTM** inclusion for **long-range dependencies**.
  - ✓ **4** stacked **ST-GCN blocks** with 256 channels followed by a **2-layer BiLSTM** encoder.
- **Gloss Prediction**
  - ✓ Feed **each feature map** to the **ST-GCN/BiLSTM/CTC model**.
  - ✓ Add decoding scores through a **posterior fusion scheme**.
  - ✓ **CTC-LSTM** models: **non-aligned spike timings**.
  - ✓ **Spike timings synchronization** using a **guiding CTC model** [6].

Datasets & Experimental Setup

- **RWTH-PHOENIX Weather 2014T dataset [7]:**
  - ✓ **German SL** videos of broadcast weather forecasts by **9 signers** (**6F**, **3M**).
  - ✓ 8,257 sequences with a **1,066-gloss vocabulary**.
  - ✓ **Multi-signer split**, comprising 7,096 training videos, 519 validation, and 642 testing.
- **Chinese SLR dataset (CSL) [8]:**
  - ✓ Studio-quality **video** of **daily-life** communication in **Chinese SL**.
  - ✓ 100 signing sentences (**178-gloss vocabulary**), expressed by **50 signers** (**25F**, **25M**) .
  - ✓ **Each sentence** is performed by each signer **5 consecutive times** (25k clips).
  - ✓ **Signer-independent setup** (**Split I**):
    - 20k training clips (40 signers) and 5k testing ones (10 signers).
    - Allocate 5k training clips for validation.

References

[1] **Papadimitriou & Potamianos,** "Multimodal sign language recognition via temporal deformable convolutional sequence learning," *Proc. Interspeech*, **2020**.
[2] **Choutas et al.**, "Monocular expressive body regression through body driven attention," *Proc. CVPR*, **2020**.
[3] **Simon et al.**, "Hand keypoint detection in single images using multiview bootstrapping," *Proc. CVPR*, **2017**.
[4] **Amorim et al.**, "Spatial-temporal graph convolutional networks for sign language recognition," *Proc. ICANN*, **2017**.
[5] **Parelli et al.,** "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," *Proc. ECCV-W*, **2020**.
[6] **Kurata & Audhkhasi,** "Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation," *Proc. Interspeech*, **2019**.
[7] **Camgöz et al.**, "Sign language transformers: Joint end-to-end sign language recognition and translation," *Proc. ICCV*, **2019**.
[8] **Huang et al.**, "Video-based sign language recognition without temporal segmentation," *Proc. AAAI*, **2018**

Experimental results

- **Comparison** of our **proposed model** to the **literature** on the **RWTH-PHOENIX Weather 2014T dataset** (PH2014T, left) and the **Chinese SLR corpus** (CSL, right) in gloss error rate (GER, %).

Model	Feature streams	PH2014T	Model	Feature streams	CSL
SFD-SGS-SFL [4]	FF + Glosses	26.10	LS-HAN [6]	FF + H + Glosses	17.30
Bi-ST-LSTM-A [5]	H + Articulations position	24.68	DenseTCN [28]	FF	14.30
Transformer-CTC [29]	FF	24.59	CTF [30]	FF	11.20
BiLSTM-CTC [3]	FF + Glosses	24.30	Align-iOpt [31]	FF	6.10
CNN-LSTM-HMM [32]	Glosses + H/M	24.10	BiLSTM + CTC [3]	FF + Glosses	2.40
Att-TDCNN [2]	H/M + 2D skel. + Flow	23.70	SLRGAN [33]	FF + Glosses	2.10
Proposed	FF + Flow + 2D/3D Pose	21.34	TMC-BiLSTM-CTC [1]	FF + H + F + Pose	2.10
TMC-BiLSTM-CTC [1]	FF + H + F + Pose	**21.00**	**Proposed**	FF + Flow + 2D/3D Pose	**1.48**

(Appearance features based on full frame (FF), hands (H), mouth region (M), and face (F). "Glosses" refers to embeddings)

- **Proposed system on:**
  - ✓ **PH2014T** : **Outperforms** most results in the **literature**, coming **very close** to the **state-of-the-art GER** (21.34% vs 21.00%).
  - ✓ **CSL**: **Achieves** the **state-of-the-art** result, significantly **outperforming** the **best alternative** by a **30% relative GER reduction** (1.48% vs. 2.10%).
- **System evaluation** on the **RWTH-PHOENIX Weather 2014T dataset**, when various **modality combinations** are considered:
  - ✓ Network yields **competitive performance** when **all three streams** are **considered**.
  - ✓ **"ExPose" parameters** boost performance.

Feature streams	GER (%)
2D skeleton	51.10
2D skeleton + Appearance	23.16
2D skeleton + Appearance + Optical Flow	22.28
3D skeleton	53.72
3D skeleton + Appearance	23.35
3D skeleton + Appearance + Optical Flow	22.37
"ExPose" parameters (Rotation)	50.25
Rotation + Appearance + Optical Flow	22.14
Joint-position + Appearance (A)	23.03
Joint-motion+ Optical Flow (B)	23.15
Rotation + Appearance (C)	22.96
A + B	22.04
A + C	21.75
A + B + C	**21.34**

- **Ablation study:**
  - ✓ Exclusion of the **BiLSTM** encoder degrades GER from 21.34% to 22.42%.
  - ✓ Exclusion of the **ST-GCNs** degrades GER from 21.34% to 24.04%.
  - ✓ Exclusion of the **guiding method** degrades GER from 21.34% to 24.84%.

Conclusions

- **Proposed** a **deep learning** model for **CSLR** from **RGB videos:**
  - ✓ **Multiple visual representations** of the signing activity.
  - ✓ **Feature** stream **combinations** into **three ST-GCN modules**.
  - ✓ **ST-GCN/BiLSTM** based **sequence learning**.
  - ✓ **Late fusion** via a **guiding CTC approach**.
- **Investigated** the **contribution** of:
  - ✓ **3D human pose** and **shape parameterization** via the "ExPose" approach.
  - ✓ **3D skeletal joint information** inferred from detected 2D joints via OpenPose.
- Showed that **fusion** of **multiple feature** streams **benefits performance**.
- Achieved competitive performance on **RWTH-PHOENIX Weather 2014T** and the new state-of-the-art on the **Chinese SLR corpus** (Split I setup).

H.F.R.I. Hellenic Foundation for Research & Innovation