Continuous Gesture Recognition without Depth Sensor using Temporal Convolutions and Recurrence

July 4, 2017

1 Team details

- Team name: deepgesture
- Team leader name: Lionel Pigou
- Team leader address, phone number and email:
iGent Tower, Technologiepark 15
9052 Zwijnaarde
Belgium
(+32)474463059
lionel.pigou@ugent.be
- Rest of the team members: /
- Team website URL (if any): /
- Affiliation: Ghent University

2 Contribution details

- Title of the contribution: Continuous Gesture Recognition without Depth Sensor using Temporal Convolutions and Recurrence.
- Final score: 0.342236
- General method description: End-to-end deep neural network on raw RGB video pixels with temporal convolutions and bidirectional LSTM networks. The model uses 20 layers and is trained without depth images nor external data.
- References: [1]
- Representative image / diagram of the method: Figure 1

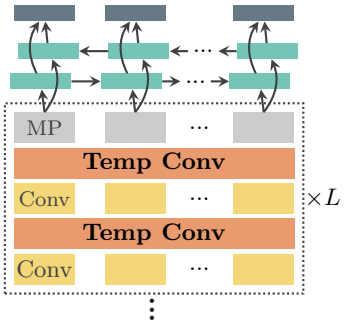

Figure 1: CNN model with bidirectional LSTM and temporal convolutions. (MP refers to max-pooling). The model has 20 layers in total.

- Describe data preprocessing techniques applied (if any): For every frame, RGB is converted to gray-scale and the preceding frame is subtracted. The depth images are not used.

3 Visual Analysis

3.1 Gesture Recognition (or/and Spotting) Stage

3.1.1 Features / Data representation

For every frame, RGB is converted to gray-scale and the preceding frame is subtracted. The depth images are not used. The frames are resized to 128x128 pixels. The raw pixels are the features that are used for the network.

3.1.2 Dimensionality reduction

No dimensionality reduction is performed other than resizing to 128x128 images.

3.1.3 Compositional model

No compositional model is used.

3.1.4 Learning strategy

The learning strategy is Adam [2] gradient descent.
To learn the model, a batch of random sequences of 32 128x128 frames is chosen. The softmax-predictions are optimized using a negative-log-likelihood loss for the 32 labels of each sequence.

3.1.5 Other techniques

The model uses residual connections [3], ELU non-linearities [4], temporal convolutions and recurrence (LSTM) [1], batch normalization [5] and data

augmentation. For evaluation, a sliding window of 32 frames is used with a stride of 16: for each 32 input frames the middle 16 predictions are used. Finally, a post-processing technique is used to smooth out predictions over the frames. The statistical *mode* over a window of 39 frames is selected for each frame.

3.1.6 Method complexity

The model uses 20 non-linearity layers and has 824,233 parameters.

3.2 Data Fusion Strategies

The network is fed by 32 input frames and outputs 32 softmax-predictions (one for each frame). The model uses temporal convolutions from the very first layer. This means that motion features are learned early in the network.

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any): /
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any): /
- Qualitative advantages of the proposed solution: The proposed solution works on RGB-only. No depth-camera is needed. Furthermore, no feature engineering is required and so the methodology can very easily be transferred to similar problems.
- Results of the comparison to other approaches (if any): /
- Novelty degree of the solution and if is has been previously published: The model is an updated version of the best model in [1].

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements):
  The model is designed, trained and evaluated using Lasagne and Theano (Python language).
  Ubuntu 16.04 LTS. 8GB GPU for training. 4GB GPU for evaluation. Data augmentation was parallelized on the CPU, but this is not required.
- Human effort required for implementation, training and validation?
  Implementation: 1 week
  Training and validation: 1 week

- Training/testing expended time?
  Video preparation: 5-8 hours
  Training: 1-2 days
  Testing: 1-2 hours

- General comments and impressions of the challenge? What do you expect from a new challenge in face and looking at people analysis?
  The organization was very professional and the data was clean to work with. I think you would attract more participants if this would be hosted on Kaggle like a few years ago.
  A new challenge focusing on sign language recognition would be a great way to advance this challenging field.

References

[1] Lionel Pigou, Aäron van den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling : recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, Oktober 2016.

[2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR 2015*, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.