Supplementary Material for "Less is More: Facial Landmarks can Recognize a Spontaneous Smile"

Md. Tahrim Faroque Tushar*[1]
tahrim.tushar@northsouth.edu
Yan Yang*[2,3]
u6169130@anu.edu.au
Md Zakir Hossain[2,3,4]
zakir.hossain@anu.edu.au
Sheikh Motahar Naim[5]
sheinaim@amazon.com
Nabeel Mohammed[1]
nabeel.mohammed@northsouth.edu
Shafin Rahman[1†]
shafin.rahman@northsouth.edu

[1] Department of Electrical and Computer Engineering, North South University, Bangladesh
[2] Biological Data Science Institute, The Australian National University, Canberra, Australia
[3] CSIRO Agriculture & Food, Canberra, Australia
[4] CSIRO Machine Learning & Artificial Intelligence Future Science Platform, Canberra, Australia
[5] Amazon Web Services

1 Dataset Details

We experiment with four smile datasets. We provide dataset statistics in Table 1 and describe them below:

Dataset	Number of Videos		Number of Subjects	
	Genuine	Posed	Genuine	Posed
UvA-NEMO	597	643	357	368
BBC	10	10	10	10
MMI	138	49	9	25
SPOS	66	14	7	7

Table 1: Statistics of smile datasets.

(a) UvA-NEMO dataset[2]. This dataset contains 597 spontaneous and 643 posed smile videos from 357 and 368 participants, respectively. It is recorded in resolution 1920×1080 at 50 frames per second. There are 400 participants distributed into 185 females and 215 males. Among them, there are 251 adults and 149 young people.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

* Equal Contribution.
† Corresponding author.

(b) BBC dataset. This dataset has 20 videos, divided into 10 posed and 10 spontaneous from 20 different subjects distributed in 6 females and 14 males. The videos are recorded in resolution 314×286 with 25 FPS.
(c) MMI facial expression dataset[4]. It contains spontaneous and posed facial expressions. Similar to [1, 6], we use 138 spontaneous and 49 posed smile videos. The spontaneous smile part is recorded in resolution 640×480 and at 29 FPS, and the posed smile videos are recorded in 720×576 pixels at 25 FPS.
(d) SPOS dataset [3]. There are two types of image sequences, gray and near-infrared sequences and all the images are in 640×480 pixels with 25 FPS. There are in total 80 smiles, 66 spontaneous, and 14 deliberate smiles.

2 Detailed Architecture

Figure 1 shows details of relativity and trajectory network of our proposed **MeshSmileNet** framework. Here, for the Relativity network, we employ CurveNet blocks, and it consists of LPFA (Local Point-Feature Aggregation) blocks and four CIC (Curve Intervention Convolution) blocks, and one Conv1d layer. In LPFA blocks, we follow the default parameters in [5] but set the maximum number of neighborhoods to 20. In the four CIC blocks, k starting points are selected via top-k selection to create curves by a learnable walk policy for aggregation, where 100 curves with a length of 5 is used. The distance between two nearby points in the curve is $0.08 - 0.08 - 0.08 - 0.16$, respectively. After the CIC blocks, we apply 1D convolution with a kernel size of 1. Then the learned feature representation propagates through the Trajectory network. The trajectory network has two transformer blocks, where 1st transformer consisting of six encoder blocks puts self-attention in the spatial domain, and the 2nd transformer with three encoder blocks puts self-attention in the time domain. Both transformer blocks use a dimension size of 256 with a dropout of 0.1. For the multi-head-attention, we use 4 heads.

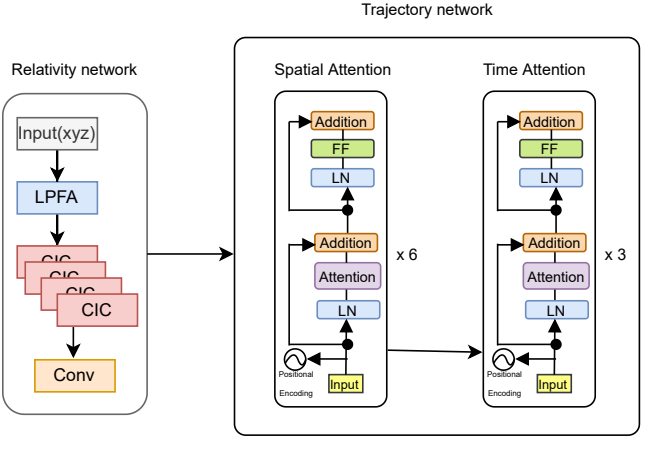

Figure 1: The detailed architecture of our proposed method: *Relativity Network* consists of one CurveNet block, and *Trajectory Network* consists of two transformer blocks with six and three encoder blocks, respectively. Here, LPFA = Local Point-Feature Aggregation, CIC = Curve Intervention Convolution, LN = Layer Normalization, FF = Feed Forward

References

[1] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012.

[2] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279–294, 2015.

[3] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 868–875. IEEE, 2011.

[4] Michel Valstar, Maja Pantic, et al. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France., 2010.

[5] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 915–924, 2021.

[6] Yan Yang, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. Realsmilenet: a deep end-to-end network for spontaneous and posed smile recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020.