# Less is More: Facial Landmarks can Recognize a Spontaneous Smile

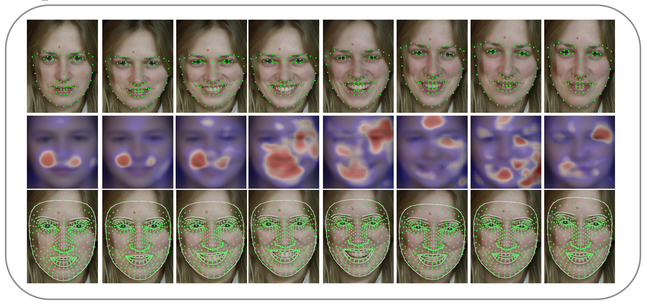
Md. Tahrim Faroque Tushar[1], Yan Yang[2,3], Md Zakir Hossain[2,3],
Sheikh Motahar Naim[4], Nabeel Mohammed[1], and Shafin Rahman[1]

[1]North South University, Bangladesh, [2]Australian National University, Australia,
[3]CSIRO, Australia, [4]Amazon Web Services, USA

BMVC 2022

## Problem Definition and Contribution

**Goal:** Establish discriminative features from facial landmarks in an end-to-end manner by considering the relativity and trajectory of the landmarks to recognize spontaneous smiles.



**Motivations:**

- Smile veracity classification is more challenging because of the difficulty of capturing and interpreting micro-facial movements in spatial and time dimensions.
- Previous work of smile veracity classification uses redundant facial features from video frames as inputs, as well as poor inductive bias and manual feature engineering which results in lower recognition performance.
- Such research aims to improve human-computer interaction and affect detection by allowing computers/robots to perceive users' emotional states.

**Key Contributions:**

- We propose a MeshSmileNet framework for classifying spontaneous and posed smiles based on facial landmarks.
- We explore the concepts, relativity and trajectory, of learning end-to-end landmark features for smile classification task.
- We perform detailed ablation studies, compare our methods with strong baselines and state-of-the-art results on four well-known smile datasets.

## Problem Formulation

**Assumption:**

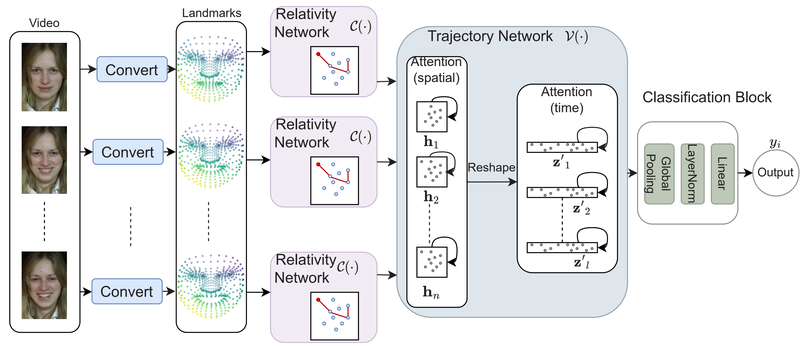
- Given a video $\mathcal{X} = [\mathbf{x}_n \mid n \in 1 \dots N]$, where $\mathbf{x}_t$ and $N$ are $t$th frame and video length, respectively, we aim to identify if $\mathcal{X}$ is a spontaneous $(y = 0)$ or posed $(y = 1)$ smile.
- Here, $y$ is the true label. From each frame, $\mathbf{x}_n \in \mathcal{X}$, we extract 3D facial landmarks, $\mathbf{a}_n \in \mathbb{R}^{3\times L}$, where $L$ is the total number of landmarks. Suppose, $\mathcal{A} = [\mathbf{a}_n \mid n \in 1 \dots N]$ denotes the landmark representation of a smile video. Our goal is to train a model $\mathcal{F}(\cdot)$ that maps $\mathcal{A}$ to $y$.

**Main Idea:**

- First, we extract $L$ number of facial landmarks $(\mathbf{a}_n)$ from each frame $(\mathbf{x}_n)$ of a video using Attention Mesh. Then, all extracted landmarks are fed into our proposed network, $\mathcal{F}(\cdot)$.
- At inference, after extracting landmarks locations representing a video, $\mathcal{A}^*$, we perform a forward pass for prediction, $\hat{y}$, i.e., $\hat{y} = \mathcal{F}(\mathcal{A}^*)$.

## Method

**Network Architecture:** Our method composes of three main components, namely (1) Relativity network, (2) Trajectory network, and (3) Classification network. In the Relativity network, We explore spatial geometry relations of landmarks at each frame ($\mathbf{a}_n \in \mathcal{A}$) based on CurveNet blocks. Then in Trajectory network, we track the movement of each landmark $\mathbf{a}_{n,l}$, $l$th landmark at $n$th frame. Finally, We classify the video into label y, spontaneous smiles or posed smiles, from the relativity and trajectory based landmark features.

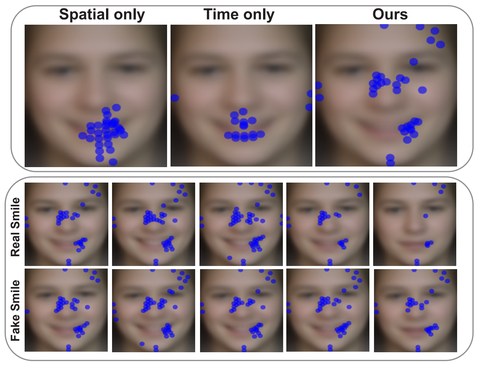



**Loss Function:** We penalize deviation of network predictions $\mathcal{F}(\mathcal{A})$ from the ground truth $y_i$ using Binary Cross Entropy loss.

$$\mathcal{L}_{BCE} = y_i \log \mathcal{F}(\mathcal{A}) + (1 - y_i) \log(1 - \mathcal{F}(\mathcal{A})) \quad (1)$$

## Analysis

- Here we showed complex interplay of relativity and trajectory networks by visualizing the importance of different landmark locations by calculating their gradient with respect to the target label. We average the background face, landmark position, and gradient of the entire UVA-NEMO dataset.
- Using self-attention across both spatial and time dimensions, our final recommendation in 'Ours' combines individual contributions (of spatial and time) and additionally focuses on the nose and eyes regions.




## Experiments & Results

**Datasets:** We experiment with four smile datasets. They are UvA-NEMO dataset, BBC dataset, MMI facial expression dataset, and SPOS dataset. We provide dataset statistics.

| 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 |

**Experimental Setups:** We present a comparison between methods (left), as well as using the same landmark extractor on different models (upper right) and experiment among diverse groups (lower right).

| Method | UVA-NEMO | MMI | SPOS | BBC |
|---|---|---|---|---|
| Cohn'04 [1] | 77.3 | 81.0 | 73.0 | 75.0 |
| Dibeklioglu'10 [2] | 71.1 | 74.0 | 68.0 | 85.0 |
| Pfister'11 [3] | 73.1 | 81.0 | 67.5 | 70.0 |
| Wu'14 [4] | 91.4 | 86.1 | 79.5 | 90.0 |
| Dibeklioglu'15 [5] | 89.8 | 88.1 | 77.5 | 90.0 |
| Wu'17 [7] | **93.9** | 92.2 | 81.2 | 90.0 |
| Mandal'17 [8] | 80.4 | - | - | - |
| Mandal'16 [6] | 78.1 | - | - | - |
| RealSmileNet'20 [9] | 82.1 | 92.0 | 86.2 | 90.0 |
| PSTNet [10] | 72.9 | 94.3 | 87.1 | 95.0 |
| P4Transformer [11] | 74.9 | 91.3 | 82.9 | 85.0 |
| Vanilla ViT [12] | 78.4 | 99.0 | 93.5 | 95.0 |
| Ours | 85.0 | **99.0** | **94.4** | **95.0** |

| Method | UVA-NEMO | MMI | SPOS | BBC |
|---|---|---|---|---|
| Dibeklioglu'12 [13] | 72.1 | 78.2 | 53.2 | 55.0 |
| Dibeklioglu'15 [5] | 77.0 | 73.9 | 46.9 | 55.0 |
| Ours | **85.0** | **99.0** | **94.4** | **95.0** |

| Individual Group UVA-NEMO | Real SmileNet [9] | Ours |
|---|---|---|
| Young | 79.6 | **80.4** |
| Adult | 79.4 | **82.4** |
| Male | 77.8 | **81.4** |
| Female | 80.0 | **80.2** |

**Ablation Studies & Significance Tests:** Different ablation studies were done based on the model's architecture (upper left), frame rates (upper right), different landmark extractor (lower right), and significance test was done between our model and baseline models.

| Input | CurveNet | Attention | UVA-NEMO | MMI | SPOS | BBC |
|---|---|---|---|---|---|---|
| Frames | No | Both* | 78.4 | 99.0 | 93.5 | 95.0 |
| Landmarks | No | Both* | 82.2 | 96.7 | 92.2 | 95.0 |
| Landmarks | Yes | Spatial | 72.3 | 98.5 | 90.1 | 95.0 |
| Landmarks | Yes | Time | 82.4 | 98.5 | 90.9 | 95.0 |
| Landmarks | Yes | Both* | **85.0** | **99.0** | **94.4** | **95.0** |

| Frame Rate | UVA-NEMO | MMI | SPOS | BBC |
|---|---|---|---|---|
| 1 | 67.0 | 98.6 | **94.4** | **95.0** |
| 3 | 74.5 | 98.6 | 90.3 | **95.0** |
| 5 | **85.0** | **99.0** | 92.4 | **95.0** |
| 10 | 82.2 | 97.6 | 92.4 | 90.0 |

| | Ours vs. ViT | Ours vs. PSTNet | Ours vs. P4Transformer |
|---|---|---|---|
| t-statistic | 8.87 | 22.97 | 10.38 |
| p-value | $9.6 \times 10^{-6}$ | $7.5 \times 10^{-8}$ | $2.6 \times 10^{-6}$ |

| Method | PSTNet | P4Transformer | Ours |
|---|---|---|---|
| DLIB [14] | 69.7 | 62.9 | 80.3 |
| Attention Mesh [15] | **72.9** | **74.9** | **85.0** |

**Acknowledgement:** Thanks to Machine Learning & Artificial Intelligence Future Science Platforms (MLAI-FSP), CSIRO for funding support.

**References:**

[1] Cohn et al. The timing of facial motion in posed and spontaneous smiles.IJWMIP (2004)
[2] Dibeklioglu et al. Eyes do not lie: Spontaneous versus posed smiles. ACM Multimedia (2010)
[3] Pfister et al. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework.ICCV Workshops (2011)
[4] Wu et al. Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors. ICASSP (2014)
[5] Dibeklioglu et al. Recognition of genuine smiles. IEEE Transactions on Multimedia, (2015)
[6] Mandal et al. Distinguishing posed and spontaneous smiles by facial dynamics. ACCV, (2016)
[7] Wu et al. Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics. ITEE, (2017)
[8] Mandal et al. Spontaneous versus posed smiles—can we tell the difference? ICCVIP, (2017)
[9] Yang et al. Realsmilenet: a deep end-to-end network for spontaneous and posed smile recognition. ACCV, (2020)
[10] Fan et al. Pstnet: Point spatio-temporal convolution on point cloud sequences. arXiv, (2022)
[11] Fan et al. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. IEEE/CVF CCVPR, (2021)
[12] Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. (2020)
[13] Dibeklioglu et al. Are you really smiling at me? spontaneous versus posed enjoyment smiles. ECCV, (2012)

Code