Customizable Facial Gesture Recognition for Improved Assistive Technology

Kuan-Chieh Wang[†‡], Jixuan Wang[†‡], Khai Truong[†], Richard Zemel[†‡]

[†]University of Toronto [‡]Vector Institute

Introduction

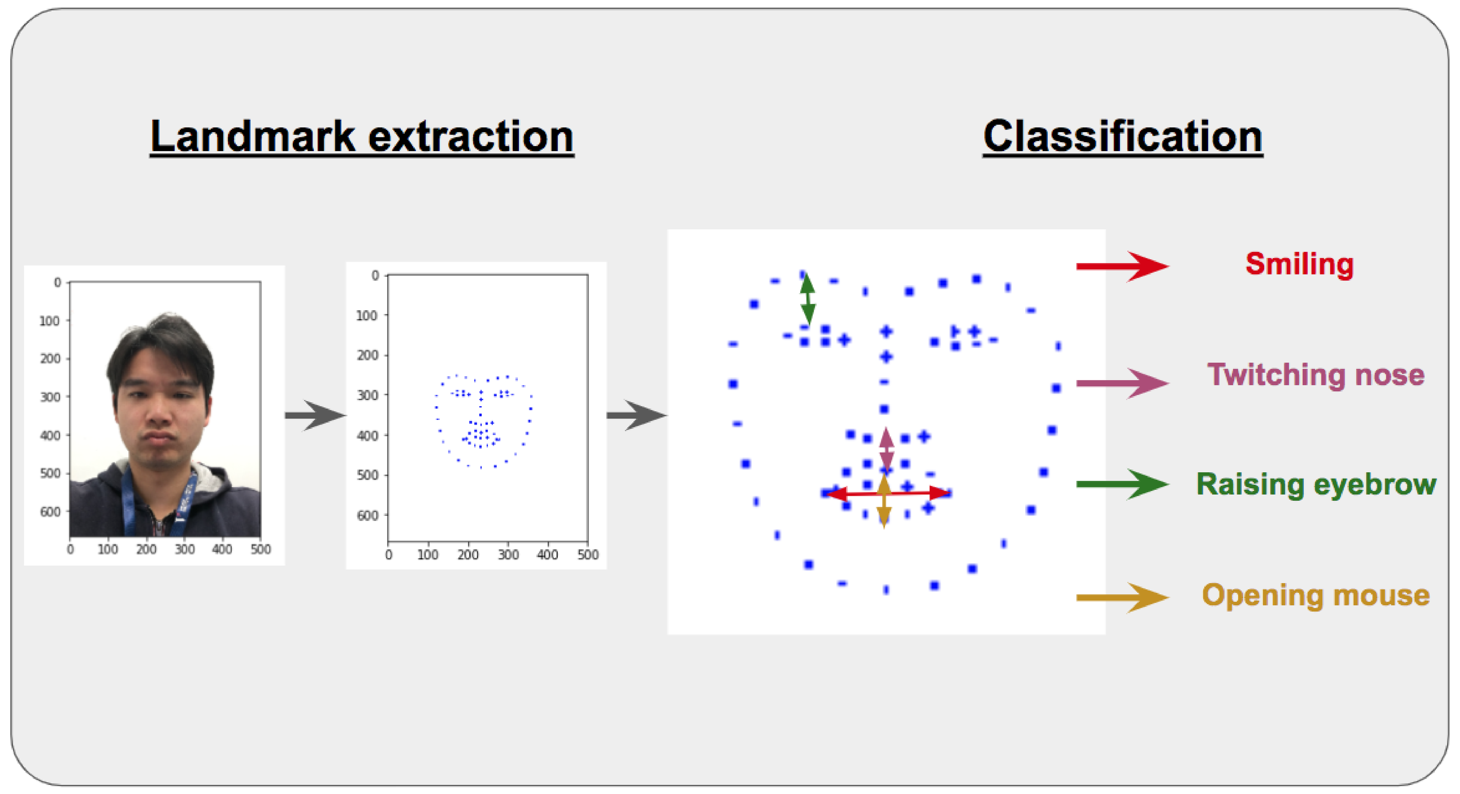
Motivation: Assistive technology based on **facial gestures** enables individuals with upper limb motor disability to interact with electronic interfaces effectively and efficiently.

Contributions:

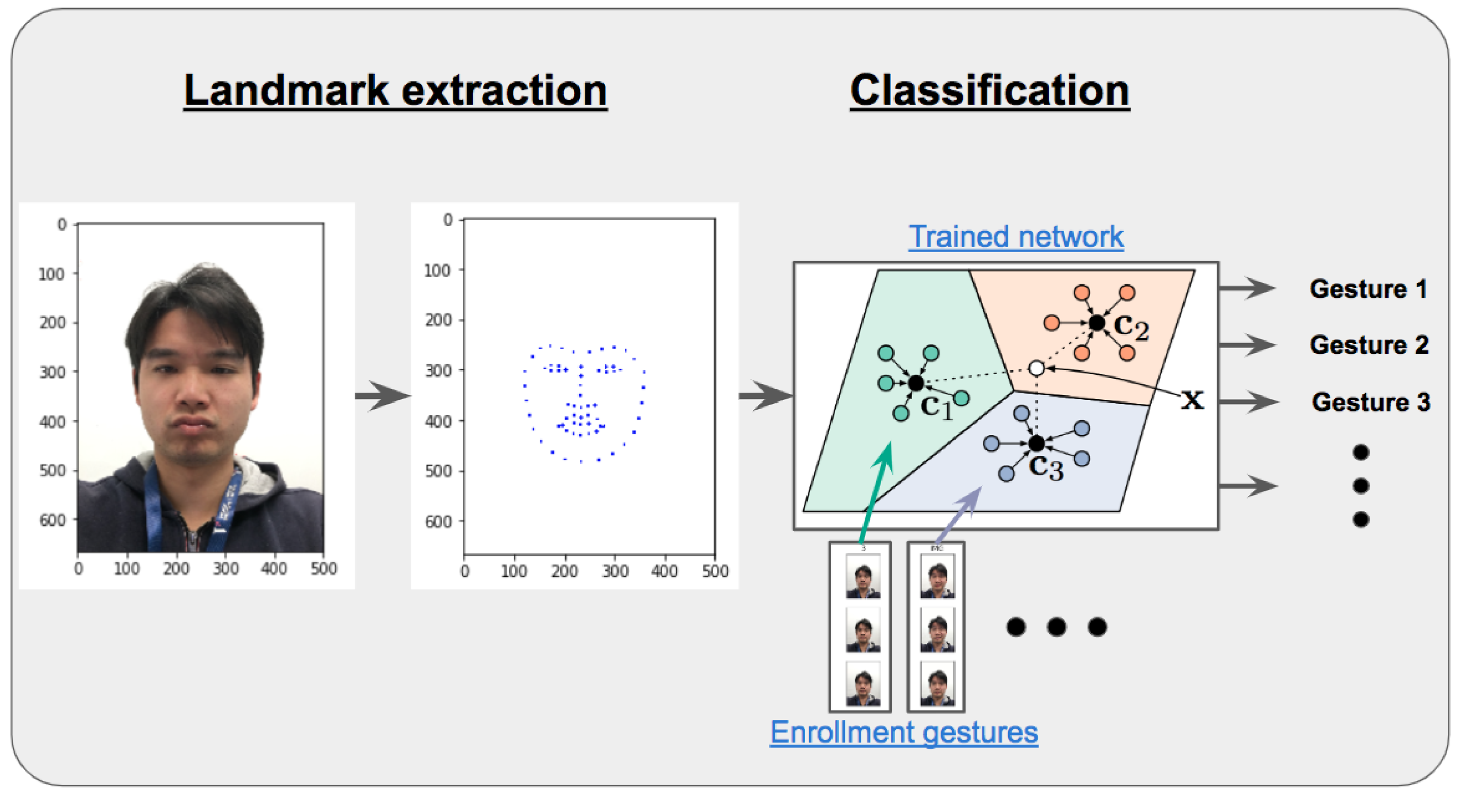
- Allows for customization by using Prototypical Networks which takes enrollment images
- Utilizes graphic engine for synthesizing training data for Prototypical Network, circumventing the need of curating a large training set manually.

Previous Work: FaceSwitch [1]

- Threshold based classifier for 4 **predefined** actions



Our modified classifier



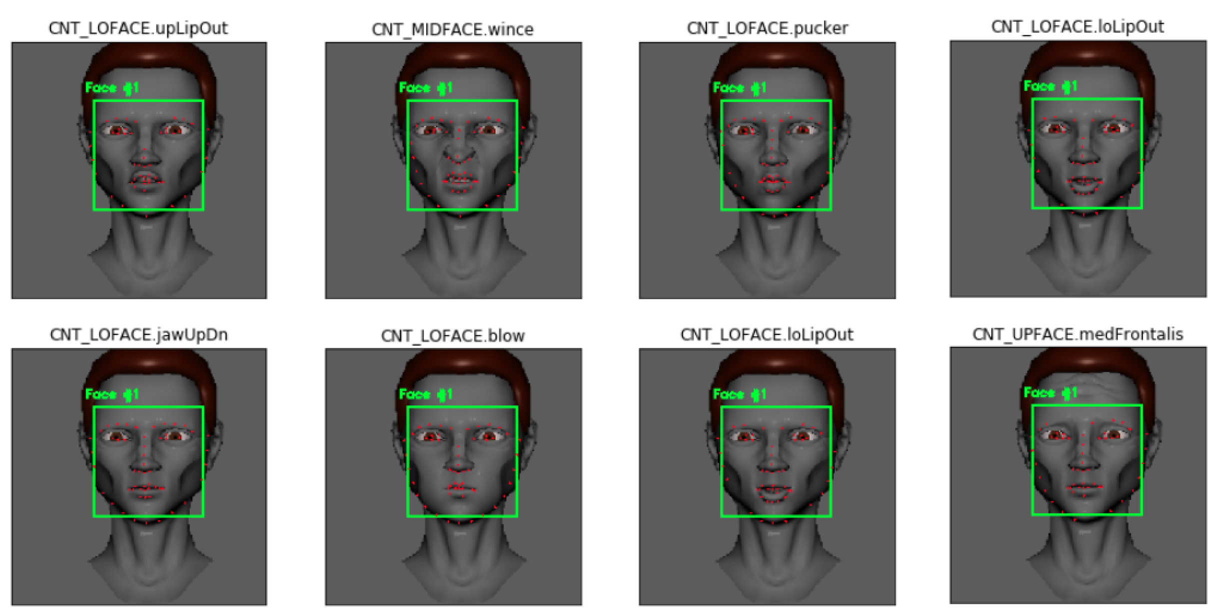
Prototypical Network

- **Support set (Enrollment images)**: $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_S}$, where $\mathbf{x}_i$ are the support images, y_i being their corresponding labels, and N_S the total number of supports.
- **Query set**: $Q = \{\mathbf{q}_i\}_{i=1}^{N_Q}$, are images to be classified into one of the support classes.
- **Prototypical Network** consists of a neural network f_ϕ, and a distance measure (e.g., Euclidean distance) $d(\cdot,\cdot)$ on the output of f_ϕ.
- A query $\mathbf{q}$ is classified based on how close it is to the class *prototype* $\boldsymbol{\mu}_c$ of each class c (computed as the average of $f_\phi(\mathbf{x})$ for all $\mathbf{x}$ in the support set S_c of class c):

$$p_\phi(y = c|\mathbf{q}) = \frac{\exp(-d(f_\phi(\mathbf{q}), \boldsymbol{\mu}_c))}{\sum_{c'} \exp(-d(f_\phi(\mathbf{q}), \boldsymbol{\mu}_{c'}))}. \quad (1)$$

Training Prototypical Network

- Successful training of few-shot classifier requires a **large training set**. E.g., the popular benchmark, Omniglot dataset, has only 20 images per class, but >**1000 classes**.
- Our insight is that, since the input are tracked landmarks of the face, we can **synthesize** a training set using a graphic engine, i.e., AutoDesk Maya.



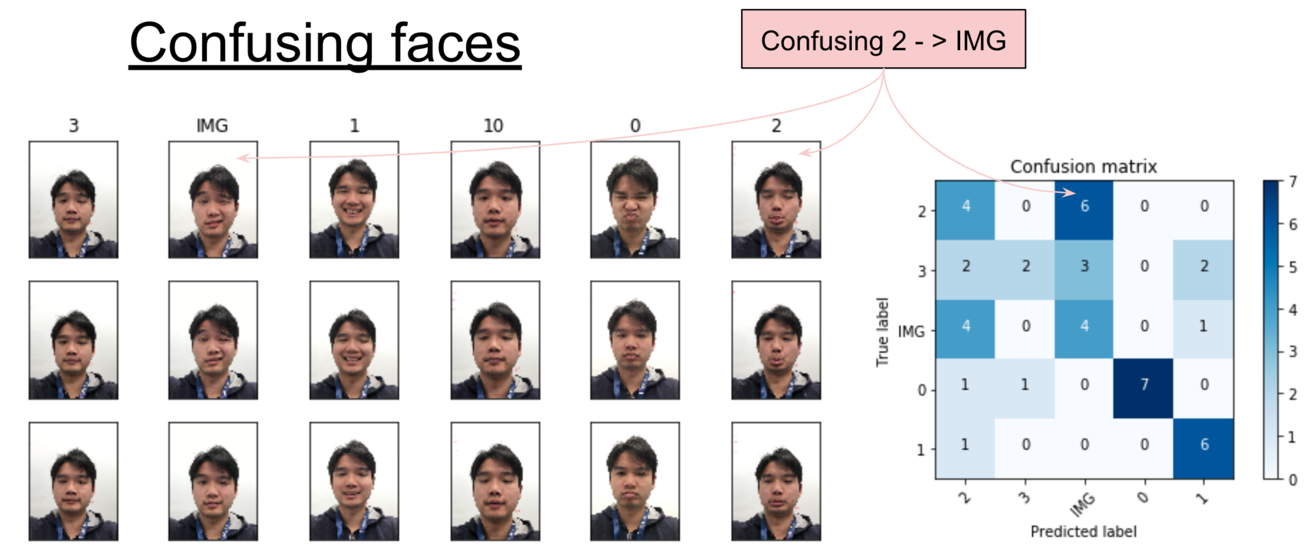
- We used the rig provided by the JALI project [2], and manually selected 15 distinct attributes.
- 225 classes were created by randomly turning on 2 of the 15 selected attributes fully, 20 samples were generated for each class.

Results

Training Setup		Accuracy on **Maya faces**		Accuracy on **Real faces**, 3-way (mean, std)		
N-way	k-shot	Train	Val (5-way, 3-shot)	1-shot	3-shot	5-shot
3	1	72.5	87.1		-	
	5	94.0	93.3	73, 12	82, 9	90, 6
	10	98.6	95.2		-	
5	1	78.0	96.9		-	
	5	98.4	92.3	57, 6	69, 14	78, 9
	10	93.2	92.1		-	
10	1	82.0	99.3		-	
	5	96.8	97.3	66, 9	67, 11	66, 8
	10	96.6	98.4		-	
50	1	86.6	99.5		-	
	5	91.4	99.8	79, 6	82, 4	85, 5
	10	94.4	98.9		-	

Table: Classification accuracy on synthesized and real faces. Evaluation on the real faces was done in the 3-way setup, using 5-shot trained models. The results on the real faces were from 3 trial runs.

- Models trained on synthetic faces can transfer to classifying real faces.



Conclusion

We present a novel method that allows AT based on facial gestures recognition to be **customizable**, and only can be trained using only **synthetic data**.

Future:

- Scale up using more diverse synthetic faces.
- Allow interaction during enrollment.

References

[1] David Rozado, Jason Niu, and Martin Lochner. Fast human-computer interaction by combining gazepointing and face gestures.ACM Trans. Access. Comput., 10(3):10:110:18, August 2017. ISSN1936-7228. doi: 10.1145/3075301. URLhttp://doi.acm.org.myaccess.library.utoronto.ca/10.1145/3075301.

[2] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization.ACM Transactions on Graphics (TOG), 35(4):127,2016.