CUSTOMIZABLE FACIAL GESTURE RECOGNITION FOR IMPROVED ASSISTIVE TECHNOLOGY

Kuan-Chieh Wang, Jixuan Wang, Khai Truong, Richard Zemel
Department of Computer Science
University of Toronto
214 College St, Toronto, ON, Canada
{wangkua1,jixuan,khai,zemel}@cs.toronto.edu

ABSTRACT

Digital devices have become an essential part of modern life. However, it is much more difficult for less able-bodied individuals to interact with them. Assistive technology based on facial gestures could potentially enable people with upper limb motor disability to interact with electronic interfaces effectively and efficiently. Previous studies proposed solution that can classify predefined facial gestures. In this study, we build a customizable facial gesture recognition system using the Prototypical Network, an effective solution to the few-shot learning problem. Our second contribution is the insight that since facial gesture recognition is done based on tracked landmarks, a training set can be synthesized using a graphics engine. We show that our model trained using only synthetic faces can perform reasonably well on realistic faces.

1 INTRODUCTION

Although electronic devices make daily activities more convenient for the able-bodied population, these benefits are rarely experienced by individuals who suffer from upper limb motor disability due to the difficulty to interact with standard interfaces. Assistive technologies (AT) that facilitate interactions between the motor-impaired individuals and computers has been a long-standing research problem in the human-computer interaction community (Istance et al., 1996).

With the rapid development of computer vision technologies, facial gestures based interaction becomes promising. Recently, Rozado et al. (2017) showed the effectiveness of their open source facial gestures based accessibility software: the FaceSwitch.

FaceSwitch combines using a eye-gaze tracker for moving the mouse and mapping facial gestures into discrete actions (e.g., mouse click and user selected keystrokes). The gaze-tracking is implemented using the combination of the Tobii X2-30 eye tracker and its companion Control Software. The contribution of Rozado et al. (2017), other than the final integrated system, is their proposed facial gesture recognition algorithm. They use the Beyond Reality Face Nxt tracker (https://www.beyond-reality-face.com/) for real-time facial landmark tracking. The outputs of the tracker are 68 2-dimensional points. Four pairs of points are chosen for each sample, and the detection of each of the four actions is done by comparing the distance between each tracked pairs to a manually decided threshold (See Figure 1).

For the 3-way classification task (i.e., 'raising eye brow','twitching nose' and 'opening mouth'), FaceSwitch is roughly 79% accurate, while the accuracy dropped to the high 60's for classification on the four gestures. For detailed performance analysis, we refer the reader to Rozado et al. (2017).

Customization The motivation for the present study is that not all targeted users would be equally capable of making the same predefined facial gestures. A more user-friendly approach should allow users to specify what gestures they would like to use through only a few enrollment images.

This study focuses on the problem of facial gesture recognition. To allow for customization, we modify the Prototypical Network (PN) to learn a embedding where facial gesture classification can be achieved using only a few enrollment examples at test time (Snell et al., 2017). Our second

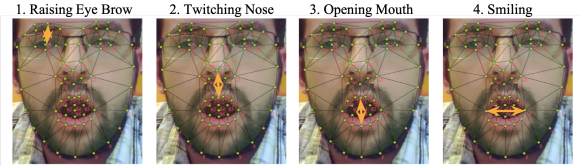

Figure 1: Illustration of the threshold based facial gesture detector in Rozado et al. (2017). Each of the double-ended arrows points to the pair of tracked points used for making decisions for one of the four predefined facial gestures.

contribution is the use of the AutoDesk Maya software for synthesizing our training set, eliminating the need for the tedious and time consuming process of data collection and annotation. The resulting system, trained on only synthetic faces, can classify real faces with reasonable accuracy. The results of this study suggest that using data generated from a graphics software can train a deep embedding model useful for gesture classification on real faces.

2 Background

Prototypical Networks (PN) is an effective algorithm for solving the few-shot classification problem (Snell et al., 2017). The few-shot problem was motivated by the phenomenon that humans can usually learn about new concepts quickly through only interacting with a handful of examples (Lake et al., 2015). At test-time, a trained model is given a small *support set*, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_S \cdot N_C}$ of new concepts (e.g., unseen classes) to learn from. Here N_S and N_C denote the number of supports per class and the number of classes, respectively. They are also commonly referred to as the 'shot' and 'way', respectively. Each $\mathbf{x}$ represents an input vector and $y_i \in \{1, ..., k\}$ is the class label for $\mathbf{x}_i$. The performance is evaluated on a *query set* $Q = \{\mathbf{q}_i\}_{i=1}^{N_Q}$. We will refer to each of the (learning S + evaluation Q) stage as one *episode* E (i.e., $E = (S, Q)$). In standard benchmarks, models are often evaluated on many episodes $\{E^j\}_{j=1}^{M}$ at test time.

For our purposes, a deployed model should be capable of quickly learning facial gestures of a new user with a few enrollment images. In the few-shot jargon, this means that a new user can customize the model by providing a support set of facial gestures.

A popular approach to train such model is through *episodic training*. Intuitively, this is a training procedure that mimicks the way of how the model will be evaluated at test-time. Using PN, a query $\mathbf{q}$ is classified based on how close it is to the class *prototype* $\boldsymbol{\mu}_c$ of class c:

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

$$\boldsymbol{\mu}_c = \frac{1}{|S_c|} \sum_{\mathbf{x}_i \in S_c} f_\phi(\mathbf{x}_i), \tag{2}$$

where $d(\cdot, \cdot)$ is a distance function such as the Euclidean distance, S_c is the support set of class c, and f_ϕ is a neural network with learnable parameters ϕ. The PN loss per episode is

$$L_{PN} = -\sum_{(\mathbf{q},c) \in Q} \log p_\phi(y = c|\mathbf{q}) \tag{3}$$

Algorithm 1 is a simple description of *episodic training*, where $\mathbb{C}_{train}$ denotes classes in the meta-train set, $\mathcal{D}_X$ the set of training examples in the set of classes in X. $\textsc{RandomSample}(s, n)$ randomly selects n elements from the set s.

One important prerequisite that makes *episodic training* successful is the availability of a large training set. This is sometimes overlooked due to the name and description of few-shot learning. For example, the most popular benchmark dataset, the Omniglot dataset, has a total of more than one thousand classes of different handwritten characters. Hence, applying the technique of few-shot learning or specifically using PN for a new problem can be ineffective due to the lack of a large training set (i.e., few examples per class, but many classes). We will discuss how this problem is addressed for our task in Section 3.

Algorithm 1 Episodic training

```
while not converged do
    $V \leftarrow \text{RANDOMSAMPLE}(\mathbb{C}_{train}, N_C)$        ▷ randomly select classes
    for $c$ in $V$ do                                                ▷ for each class
        $S_c \leftarrow \text{RANDOMSAMPLE}(\mathcal{D}_{\{c\}}, N_S)$   ▷ select the support set
        $Q_c \leftarrow \text{RANDOMSAMPLE}(\mathcal{D}_{\{c\}} \setminus S_c, N_Q)$   ▷ select the query set
    end for
    $\phi \leftarrow \phi - \alpha \nabla L_{PN}$
end while
```

3 Method

Our insight is that, since tracked landmarks were shown to be useful and robust representation for classification of facial gestures in Rozado et al. (2017), they would also be good candidate as inputs for the PN. Since the tracked landmarks are invariant to most variations in a natural facial image (e.g., lighting, skin tone, texture and so on), we can replace the laborious effort of collecting a training set of real faces with synthesizing faces of different gestures using a graphics software.

Synthetic Training Set For synthesizing training faces with different expressions, we use the rig provided by the JALI project based on the AutoDesk Maya software (Edwards et al., 2016). A rig is a model of an animated object that can be easily controlled by a handful of attributes. The rig from JALI has attributes such as 'grimace', 'pucker', and squinting each of the eyes. We manually went through all the attributes and selected 15 that do not correspond to similar gestures to encourage diversity in our samples (See Figure 2).

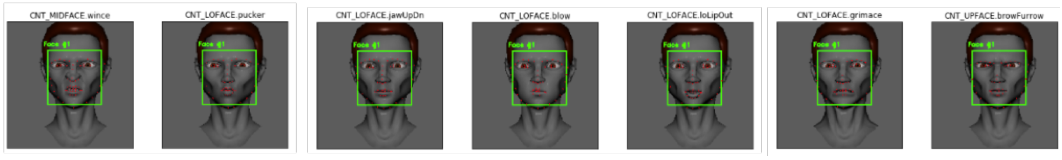

Figure 2: Examples of faces of which each has one of the selected attributes turned on.

To generate our training set, we created 225 classes by randomly turning on 2 of the 15 selected attributes fully. For samples in each class, we perturb the off attributes by roughly 10% using noise from a uniform distribution to account for the diversity of real faces. 20 samples are generated for each class, resulting in a training set of 4500 examples.

Modified PN The key modification to PN is to use the landmark features as input $\mathbf{x}$. [1] Since now $\mathbf{x}$ is a 136-dimensional vector instead of an image, we replace the original ConvNet architecture with a fully-connected MLP with 3 hidden layers of 64 units. We also use BatchNorm (Ioffe & Szegedy, 2015) and ReLU activation (Nair & Hinton, 2010).

4 Experiments

As observed in Snell et al. (2017), using different numbers of way and shot during training can affect the performance at test time. The heuristics is to use a slightly higher train way and same shot as the test episodes. Since the suggested mode of operation in Rozado et al. (2017) is to turn on one of the 3 gestures at a time, we evaluate our model on the real faces in the 3-way setting.

Table 1 shows results on both the synthetic Maya faces and real faces. The synthetic dataset is split into a training set of 200 training classes and 25 validation classes. Using more shots during training naturally increases the training accuracy. However, there is no observed trend among the validation performances. Using a higher way in training improves validation performance.

[1] we used the Python version of `http://dlib.net` for landmark extraction.

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 1: Classification accuracy of our method. Evaluation on the real faces are done in the 3-way setup. Evaluation on the real faces are only done on the 5-shot trained models. The results on the real faces are from 3 trial runs.

Then, we choose models trained in the 5-shot setting to evaluate on the real faces that were collected manually. The real faces dataset was prepared by taking ten photos of each of the three expressions from one human subject. Between repetition of the same expression, the subject is asked reset to a neutral face. Increasing the shot of real faces increases the accuracy, but this comes with the tradeoff that users would need to register more enrollment images. Our model trained on the 50-way setting can achieve an classification accuracy of roughly 80% on real faces even when only one enrollment image is provided and reach 85% with five enrollment images.

5 Related Work

Facial gesture recognition Back in 1978, the Facial Action Coding System (FACS) defined representation of facial expression by specific action units (AU) and their temporal segments that produced the expression (Friesen & Ekman, 1978). Since then, researchers have studied better representation for facial expression recognition (De la Torre & Cohn, 2011; Martinez & Du, 2012; Valstar et al., 2012). Many existing works were based on laboratory datasets and perform classification on a prefixed set of expressions and AU (Chu et al., 2013; Chen et al., 2013). Little work has been done that focuses on personalized systems in realistic scenarios. Zen et al. (2016) presented a transfer learning approach that could learn the mapping between person-specific sample distributions and between parameters of corresponding classifiers. Our approach focuses on allowing users to customize by utilizing the Prototypical Network (Snell et al., 2017).

ATs based on facial gestures Various ATs based on facial gestures were created for motor disabilities (Kouroupetroglou, 2013). Astler et al. (2011) utilized facial gesture recognition to relay nonverbal messages to blind individuals. Karpov et al. (2011) proposed a bi-modal interface using both speech recognition and facial gesture recognition. Wang et al. (2018) created a wearable interface for motor disabled individuals to play games. Our work utilizes the same facial landmark as FaceSwitch, which is an open source software to help motor-impaired people interact with computers (Rozado et al., 2017).

Sim2Real Neural Networks Lastly, one of our contributions is to utilize a graphics engine for generating the training set whereas the final model is tested on realistic images. The idea of using a rendering engine for tasks based on realistic images was shown to be very effective in Shrivastava et al. (2017), where they used a rendering engine to synthesize images of the eye to improve eye gaze regression on realistic eye images.

6 Conclusion

In this study, we present a novel method that allows AT based on facial gestures recognition to be customizable. This is achieved using a modified Protoypical Network that takes in tracked landmarks as inputs. Our second contribution is the insight that since the classification is done based on tracked landmarks, a large training set can be generated using a graphics engine. Finally, we show that our model trained only on synthetic images can perform reasonably well on realistic faces.

References

Douglas Astler, Harrison Chau, Kailin Hsu, Alvin Hua, Andrew Kannan, Lydia Lei, Melissa Nathanson, Esmaeel Paryavi, Michelle Rosen, Hayato Unno, et al. Increased accessibility to nonverbal communication through facial and expression recognition technologies for blind/visually impaired subjects. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pp. 259–260. ACM, 2011.

Jixu Chen, Xiaoming Liu, Peter Tu, and Amy Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15):1964–1970, 2013.

Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3515–3522, 2013.

Fernando De la Torre and Jeffrey F Cohn. Facial expression analysis. In *Visual analysis of humans*, pp. 377–409. Springer, 2011.

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.

E Friesen and P Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Howell Owen Istance, Christian Spinner, and Peter Alan Howarth. Providing motor impaired users with access to standard graphical user interface (gui) software via eye-based interaction. 1996.

Alexey Karpov, Andrey Ronzhin, and Irina Kipyatkova. An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. In *International Conference on Human-Computer Interaction*, pp. 454–463. Springer, 2011.

Georgios Kouroupetroglou. *Assistive technologies and computer access for motor disabilities*. IGI Global, 2013.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Aleix Martinez and Shichuan Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research*, 13(May): 1589–1608, 2012.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

David Rozado, Jason Niu, and Martin Lochner. Fast human-computer interaction by combining gaze pointing and face gestures. *ACM Trans. Access. Comput.*, 10(3):10:1–10:18, August 2017. ISSN 1936-7228. doi: 10.1145/3075301. URL `http://doi.acm.org.myaccess.library.utoronto.ca/10.1145/3075301`.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107–2116, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):966–979, 2012.

Ker-Jiun Wang, Quanbo Liu, Yifan Zhao, Caroline Yan Zheng, Soumya Vhasure, Quanfeng Liu, Prakash Thakur, Mingui Sun, and Zhi-Hong Mao. Intelligent wearable virtual reality (vr) gaming controller for people with motor disabilities. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 161–164. IEEE, 2018.

Gloria Zen, Lorenzo Porzi, Enver Sangineto, Elisa Ricci, and Nicu Sebe. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*, 18(4):775–788, 2016.