A Sim2Real Approach to Augment Low-Resource Data for Dynamic Emotion Expression Recognition

Saba Akhyani Mehryar Abbasi Mo Chen Angelica Lim
Simon Fraser University
sakhyani@sfu.ca, mabbasib@sfu.ca, mochen@cs.sfu.ca, angelica@sfu.ca

Abstract

Robots and artificial agents that interact with humans should be able to do so without bias and inequity, but facial perception systems have notoriously been found to work more poorly for certain groups than others. In our work, we aim to build systems that can perceive humans in a more transparent and inclusive manner. Specifically, we focus on dynamic expressions on the human face, which are difficult to collect for a broad set of people due to privacy concerns and the fact that faces are inherently identifiable. Furthermore, datasets collected from the Internet are not necessarily representative of the general population. We address this problem by 1) creating a suite of 3D simulated human models to generate complex facial expressions outside of the prototypical basic emotions, such as confusion, 2) creating an auditable synthetic dataset with pre-defined coverage over given ethnic and gender parameters, and 3) covering the varied viewing angles a robot may encounter a human in the real world. The addition of synthetic training data to the proposed model architecture showed an accuracy improvement of 11% on our dynamic facial expression recognition task, compared to the performance of the same model architecture without synthetic training data. We also show that this additional step reduces the adverse effect of dataset bias when the architecture's feature extraction weights are trained from scratch.

1. INTRODUCTION

A wide range of human-robot interaction (HRI) applications can benefit from social robots that perceive social signals and react accordingly. For instance, imagine a restaurant service robot seeing a look of confusion wash over your face as you look at the menu. It detects your hesitation and proactively offers assistance: "Do you have any questions I can answer?". This simple scenario plays out between humans each day all over the world, but robots are still far from capable of performing this kind of proactive assistance in a robust manner. Several major challenges prevent such systems from being deployed in the real world.

The first challenge is a lack of data for dynamic facial expressions outside of the prototypical set of emotions [3]. In recent years with the advent of deep learning methods, efforts have been made for improvements towards facial emotion recognition (FER) [17, 32, 35, 21]. While these classifiers are increasing in accuracy on static image databases, it is not clear they can effectively estimate emotions in video streams [3] as human behavior is dynamic in nature [15, 7] (e.g. looking left and right to indicate confusion). Furthermore, in HRI, as stated in a review by [25], "there are a wide range of possible affective levels expressed by people in HRI that the robot needs to understand in order to participate in a bi-directional social interaction with humans.". Most facial expression datasets focus on a specific set of emotions such as happiness, sadness, anger, surprise, fear, disgust, and neutral [8]. Therefore, in this work, we address the challenge of classifying *dynamic* expressions, focusing here on an understudied social signal of *confusion*.

Secondly, social robots should also have the ability to evaluate human affective expressions fairly, without discriminating against underrepresented groups. A recent survey on automatic multi-modal emotion recognition in the wild shows that inclusivity of all ethnicities remains a challenge in emotion recognition systems and should be further investigated [30]. According to [27, 5], racial bias is apparent in current machine learning methods, especially those involving the face. For example, [5] found that commercial face gender classification systems all perform better on male and light-skinned faces, and are least accurate on dark-skinned females. In another study, Rhue [27] used a dataset of basketball player photos and found racial biases in Face++ and Microsoft's Face API. Black men were more likely to be tagged with a negative emotion than white men. Kara et al. [19] also demonstrated how using continual learning can lead to fairer facial emotion recognition and facial action unit (AU) detection algorithms. One major cause of this bias is that major FER datasets are underrepresentative of genders and racialized backgrounds [33]. As

faces are inherently identifiable, ethical and privacy concerns can be an issue if requiring real humans to provide their data [18], yet anonymizing the face can remove important facial features. To address this issue, we create the first (to our knowledge) dataset of synthetic humans that specifically incorporates multiple ethnicities to improve dynamic facial emotion expression recognition.

Finally, deep learning methods for user emotion modeling is especially challenging because collecting and labeling large amounts of naturalistic and spontaneous user facial expressions is very time-consuming [31]. Facial emotional expressions are difficult to label due to the subjective nature of annotation. Therefore, such datasets, especially for videos, are usually small compared to those in domains where deep learning has been most successful, such as in object recognition. In this study, we employ a simulation to reality (Sim2Real) approach to address the previously mentioned challenges. Sim2Real approaches have performed well in different domains such as hand tracking [26] and text detection [12]. For face recognition and facial feature detection, some studies have attempted to address this problem by augmenting datasets [29, 20, 2] to include variations of pose, camera locations, and illumination conditions [10, 1, 13, 24, 23, 22]. They find that networks trained on the synthetic dataset could generalize well on real-world static images. Similarly, we generate a diverse dataset that includes **low-resource ethnicities**, **new expressions**, varied angles and lighting, and focus on creating a network that can detect the **dynamic** and **understudied** confusion social signal among those close to it, such as disgust and anger.

2. Methodology

2.1. Dataset Generation

In this section, we describe the collection of in-the-wild emotionally expressive videos, and the generation of synthetic videos using a suite of simulated humans. We focus this study on a dynamic social signal which is understudied, yet common in HRI[16]: confusion, which lacks data on the web [14]. Confusion has specific characteristics, such as eyelid tightening and/or frowning which can be mistaken for social signals of anger and disgust.

2.1.1 Collection of in-the-wild confusion, anger and disgust videos

We collected short video clips (1-3s) from YouTube.com and Giphy.com using search tags such as “angry”, “confused”, and “disgust” reactions to gather human facial expressions of our desired social signals. Each video was then labeled by two annotators identifying with Canadian culture (inter-rater agreement kappa score= .88), discarding the low confidence videos. We created a multi-ethnicity dataset of real human videos expressing the three understudied social signals of confusion (45 videos), anger (47 videos), and disgust (45 videos). The final dataset contains 137 videos, of which 26 are of non-Caucasian subjects.

2.1.2 Creation of a suite of simulated humans

The overarching vision of this work is to create a large, auditable suite of human models to represent people from many different backgrounds. As a first step towards this goal, we create 24 simulated human adult models balanced on gender and four different ethnicities (Caucasian, Black, Asian, and Hispanic).

We use the MakeHuman toolkit [4], which is an open-source and free 3D computer graphics toolset, designed for prototyping human-like models. The MassProduce plugin within the MakeHuman application was then used to create several randomly generated human models of multiple ethnicities and different ages and skin colors. Out of all those generated human-like models, we selected 24 models for our study (9 samples are shown in Fig. 1a).

(a) (b)

Figure 1: (a) Sample of generated human-like models. (b) visualizing the viewing angles.

2.1.3 Generation of augmented dynamic social signal video dataset

The task of creating desired social signals is made possible using the FACSHuman software [11]. FACSHuman offers the possibility of manipulating the Action Units (AU) presented in the Facial Action Coding System (FACS) [8] on the 3D models created in the Makehuman software.

Multiple social signals per emotion We used the FACSHuman software to create 21 different social signal animations. We focus on deepening our understanding of the understudied emotion classes, by identifying 7 sub social signals for each class. These animations convey multiple variations of social cues such as anger and disgust. These 21 social signals were manually animated over 25 frames, and were created via inspection of the in-the-wild real human dataset. For example, Fig 2 shows the AUs that were

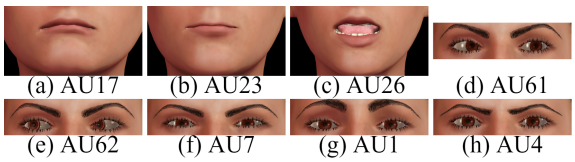

Figure 2: (a) Chin raiser, (b) Lip tightener , (c) Jaw drop, (d) Eyes left, (e) Eyes right, (f) Lid tightener, (g) Inner brow raiser, (h) Brow lowerer

used to create dynamic confusion social signals. Varied AU combinations and sequences were used to animate the 21 social signals. An example is a side-eye movement confusion state made by a timed sequence of the following AUs: AU61, AU62, AU61. While future work should perform this animation creation process automatically from video data, the manual animation creation step in this study allows us to validate the Sim2Real portion given human-level feature extraction.

Multiple viewing angles As robots may view a human from varied angles, it is important that our generated dataset incorporate varied perspectives. Our dataset was therefore expanded by creating videos of the same facial gesture but from multiple viewing angles, to make our model invariant to the face viewing angle. The camera movement included horizontal rotations of $-40, -20, 0, 20, 40$ and vertical rotations of $-30, 15, 0, 15, 30$. The nine combinations of $(H_{rotation}, V_{rotation})$=$\{(-40, -30)$, $(-20, -15)$, $(0, 0)$, $(20, 15)$, $(40, 30)$, $(40, -30)$, $(20, -15)$, $(20, -15)$, $(40, -30)\}$ were selected as our viewing angles (Fig. 1b).

2.2. Data preparation

In order to refine the data and remove any unimportant or unrelated information in the images, we used Multi-task CNN (MTCNN) to detect and crop the faces before feeding frames to our network [36] and resized images to 160×160. Additional transforms were also applied to the images randomly on each epoch, including cropping, perspective, affine transform, horizontal flip, and color transforms (shown in Fig. 3). We ensured that the same transformations were applied to all the frames from the same video.

2.3. Model Architectures

We chose to compare two deep neural network (DNN) architectures for this video classification task, and compare them to a baseline K-Nearest Neighbours classifier(KNN) useful for small datasets. This baseline model uses FaceNet [28] for frame facial feature extraction and a KNN with a Dynamic Time Warping (DTW) [34] metric as the video classifier. We refer to this baseline model as FN+KNN.

The first DNN architecture is a decoupled feature-extraction and time-series classifier approach. This architecture incorporates an image-based facial feature extraction layer from FaceNet [28], and InceptionTime128 [9] as our classifier, the latter having been proven to be a versatile and promising machine learning solution for many time series classification tasks [9]. The complete structure is shown in Fig. 4. This model is referred to here as FN+INC25 or FN+INC64. The number in the name indicates the video input number of frames. Shorter videos are padded, and the longer video are cropped out from the middle.The second DNN architecture is I3D [6], an advanced video classification method that applies combined temporal and spatial processing using 3D convolutional layers. This architecture enables us to perform evaluation using a model that is not tailored toward facial image processing.

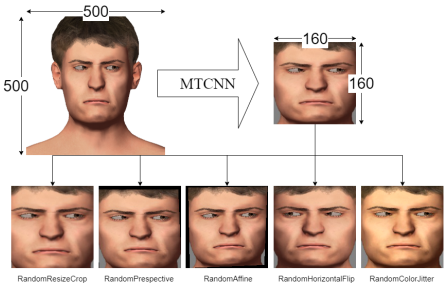

Figure 3: Image preprocessing and augmentation.

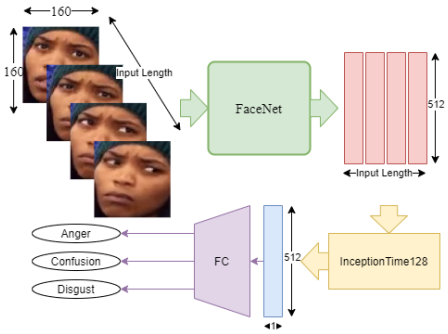

Figure 4: Model architecture for FaceNet+InceptionTime128

3. Experimental Results

We performed several experiments varying the architecture, input length and the use of only synthetic, synthetic plus real data, or only real data. We used a 5-fold cross-validation for comparison of the performances of different tests, with 1 fold consisting primarily of expressions by non-Caucasian individuals.

Training the model on synthetic (+ real) data

In this training method, the model was first trained on the synthetic dataset alone. The simulated human models were randomly divided into two sets of 19 and 5 models. All of the generated videos using simulated human models in the larger set were used for training, and those in the smaller set were used for validation. The respective number for the videos in the training and validation data were 3591 and 945. The training was done over 20 epochs with the learning rate of 10^{-4}. The batch size was set to 8. After the completion of the first step, we retrained the model on a selected training folds of the real dataset, over 50 epochs. The model is tested on the remaining single test fold. This operation is repeated 5 times each time a new fold is selected as the test fold.

The results for these experiments are shown in Table 1. These results are the average of all 5 runs. We also included experiments in which the synthetic data training step was skipped in order to highlight the effect of the addition of synthetic data training. Additionally, we compared our methods with the baseline FN+KNN classifier applied only to the real data. Our results showed that the models trained with synthetic data outperformed their counterpart only trained on the real data. The I3D model was especially impacted by the addition of synthetic data, and its accuracy of 80% outperforms all of the models that were not influenced by the synthetic data. This is quite impressive because this model was designed for video action recognition tasks. Unlike the other models, the I3D had no prior information about the facial features. The best model was the FaceNet+InceptionTime model with the input length set to 25. The confusion matrix of this model is shown in Fig. 5.

In Table 2 we explored the effect of synthetic data on the real data fold that primarily included the videos of the underrepresented (non-Caucasian) ethnicities. A fascinating result is that the I3D model has the highest increase of accuracy among all of them. However, the I3D model has higher accuracy on this fold than its average shown in Table 1. The lack of prior facial knowledge may explain its comparative lack of racial bias or performance degradation.

Network	Length	Syn[1]	Fs[2]	Acc[3]
FN+KNN	25	✗	66	67
FN+KNN	64	✗	68	69
FN+KNN	283	✗	68	69
FN+INC25	25	✗	76	76
FN+INC25	25	✓	**86**	**87**
FN+INC64	64	✗	78	78
FN+INC64	64	✓	84	84
I3D	64	✗	62	64
I3D	64	✓	79	80

[1] Synthetic Data [2] F-score [3] Accuracy

Table 1: Performance comparison of all models

Network	Syn[1]	Length	Prc[2]	Rec[3]	Fs[4]	Acc[5]
FN+INC25	✗	25	79	78	77	78
FN+INC64	✗	64	82	81	81	81
I3D	✗	**64**	**81**	**74**	**71**	**74**
FN+INC25	✓	25	84	82	81	82
FN+INC64	✓	64	82	82	81	82
I3D	✓	**64**	**86**	**82**	**81**	**82**

[1] Synthetic Data [2] Precision [3] Recall [4] F-score [5] Accuracy

Table 2: Effect of our synthetic data pre-training on correct classification within non-Caucasian data fold

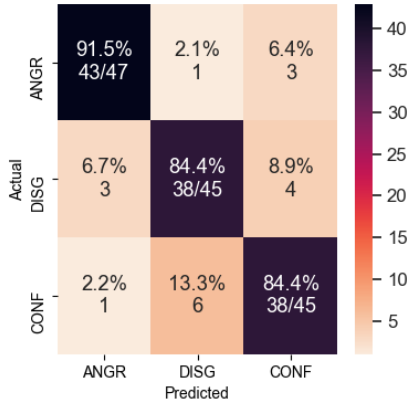

Figure 5: Combined confusion matrix of all the folds for the FN+INC25

4. Conclusions and Future Work

We showed that our Sim2Real approach improves FN+INC25 performance on our dynamic facial expression recognition task by 11% up to 87%, compared to to the performance of the same model architecture without synthetic training data. It was also shown that the addition of multi-ethnicity synthetic data reduces the adverse effect of dataset bias when the architecture's feature extraction is trained from scratch. Future work can explore other emotional expressions that are found in real human-robot interactions but lack data, as well as implementing the expression detector on a real robot with a diverse set of people. An increased number of simulated human models may also improve the overall accuracy, especially if they can be made with photo-realistic 3D model generating engines. The animation generation process from video data can be automated as the next steps of this study.

References

[1] Iman Abbasnejad, Sridha Sridharan, Dung Nguyen, Simon Denman, Clinton Fookes, and Simon Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. pages 1609–1618, 10 2017. 2

[2] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *CoRR*, abs/1809.02169, 2018. 2

[3] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. PMID: 31313636. 1

[4] Manuel Bastioni, Simone Re, and Shakti Misra. Ideas and methods for modeling 3d human figures: the principal algorithms used by makehuman and their implementation in a new approach to parametric modeling. In *Proceedings of the 1st Bangalore Annual Compute Conference*, pages 1–6, 2008. 2

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. 1

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[7] Fadi Dornaika and Bogdan Raducanu. Efficient facial expression recognition for human robot interaction. In *International Work-Conference on Artificial Neural Networks*, pages 700–708. Springer, 2007. 1

[8] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 1, 2

[9] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. 3

[10] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[11] Michaël Gilbert, Samuel Demarchi, and Isabel Urdapilleta. Facshuman a software to create experimental material by modeling 3d facial expression. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 333–334, 2018. 2

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images, 2016. 2

[13] Jian Han, Sezer Karaoglu, Hoang An Le, and T. Gevers. Improving face detection performance with 3d-rendered synthetic data. 12 2018. 2

[14] Michal Hucko, Robert Moro, and Maria Bielikova. Confusion detection dataset of mouse and eye movements. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 281–286, 2020. 2

[15] Rachael E. Jack and Philippe G. Schyns. Toward a social psychophysics of face communication. *Annual Review of Psychology*, 68(1):269–297, Jan. 2017. 1

[16] Ghazal Saheb Jam, Jimin Rhim, and Angelica Lim. Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Mar. 2021. 2

[17] Mahesh Jangid, Pranjul Paharia, and Sumit Srivastava. Video-based facial expression recognition using a deep learning approach. In *Advances in Computer Communication and Computational Sciences*, pages 653–660. Springer, 2019. 1

[18] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, July 2015. 2

[19] Ozgur Kara, Nikhil Churamani, and Hatice Gunes. Towards fair affective robotics: Continual learning for mitigating bias in facial expression and action unit recognition, 2021. 1

[20] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *CoRR*, abs/1908.04913, 2019. 2

[21] Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, and Soo-Young Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–57, 2016. 1

[22] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2093–2102, 2018. 2

[23] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. pages 2261–2268, 06 2019. 2

[24] Adam Kortylewski, A. Schneider, T. Gerig, B. Egger, Andreas Morel-Forster, and T. Vetter. Training deep face recognition systems with synthetic data. *ArXiv*, abs/1802.05891, 2018. 2

[25] Derek McColl, Alexander Hong, Naoaki Hatakeyama, Goldie Nejat, and Beno Benhabib. A survey of autonomous human affect detection methods for social robots engaged

in natural HRI. *Journal of Intelligent & Robotic Systems*, 82(1):101–133, Aug. 2015. 1

[26] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[27] Lauren Rhue. Racial influence on automated perceptions of emotions. *SSRN Electronic Journal*, 2018. 1

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *arXiv:1503.03832 [cs]*, June 2015. arXiv: 1503.03832. 3

[29] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017. 2

[30] Garima Sharma and Abhinav Dhall. A survey on automatic multimodal emotion recognition in the wild. In *Advances in Data Science: Methodologies and Applications*, pages 35–64. Springer International Publishing, Aug. 2020. 1

[31] Shane Sims and Cristina Conati. A Neural Architecture for Detecting Confusion in Eye-tracking Data. *arXiv:2003.06434 [cs, eess]*, Mar. 2020. arXiv: 2003.06434. 2

[32] Yichuan Tang. Deep learning using support vector machines. *CoRR, abs/1306.0239*, 2, 2013. 1

[33] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017. 1

[34] Byoung-Kee Yi, Hosagrahar V Jagadish, and Christos Faloutsos. Efficient retrieval of similar time sequences under time warping. In *Proceedings 14th International Conference on Data Engineering*, pages 201–208. IEEE, 1998. 3

[35] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015. 1

[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 3