DyadGAN: Generating Facial Expressions in Dyadic Interactions

Yuchi Huang Saad M. Khan
Educational Testing Service
660 Rosedale Rd, Princeton, NJ 08540
yhuang001,skhan002@ets.org

Abstract

Generative Adversarial Networks (GANs) have been shown to produce synthetic face images of compelling realism. In this work, we present a conditional GAN approach to generate contextually valid facial expressions in dyadic human interactions. In contrast to previous work employing conditions related to facial attributes of generated identities, we focused on dyads in an attempt to model the relationship and influence of one person's facial expressions in the reaction of the other. To this end, we introduced a two level optimization of GANs in interviewer-interviewee dyadic interactions. In the first stage we generate face sketches of the interviewer conditioned on facial expressions of the interviewee. The second stage synthesizes complete face images conditioned on the face sketches generated in the first stage. We demonstrated that our model is effective at generating visually compelling face images in dyadic interactions. Moreover we quantitatively showed that the facial expressions depicted in the generated interviewer face images reflect valid emotional reactions to the interviewee behavior.

1. Introduction

Advances in automated speech recognition and natural language processing have made possible virtual personal assistants such as Apple Siri, Amazon Alexa and Google Home among others. These virtual assistants have found application in a plethora of everyday activities from helping people manage daily schedules and appointments, to searching the Internet for their favorite songs. However, being primarily speech driven, such virtual agents are inherently limited in their ability to sense and understand user behavior and thereby adequately address their needs. Human interaction is a highly complex interplay of verbal and non-verbal communication patterns that among other skills demonstrates a keen ability to convey meaning through finely calibrated facial expressions [9]. Recent research in autonomous avatars [6, 22] aims to develop powerful human-computer interfaces that mimic such abilities. Not only do these avatar systems sense human behavior holistically using a multitude of sensory modalities, they also aim to embody ecologically valid human gestures, paralinguistics and facial expressions.

However, producing realistic facial expressions in avatars that are appropriately contextualized and responsive to the interacting human remains a significant challenge. Early work on facial expression synthesis [4, 5] often relied on rule based systems that mapped emotional states to predefined deformation in 2D or 3D face models. Such knowledge based systems have traditionally utilized the Facial Action Coding systems [8], which delineates a relationship between facial muscle contractions and human emotional states. Later, statistical tools such as principal component analysis were introduced to model face shapes as a linear combination of prototypical expression basis. By varying the base coefficients a shape model is optimized to fit existing images or create new facial expressions [3]. A key challenge for such approaches is that the full range of appearance variations required for convincing facial expression is far greater than the variation captured by a limited set of rules and base shapes. Advanced motion capture techniques have also been used to track facial movement of actors and transfer them to avatars [16] recreating highly realistic facial expressions. However, these solutions are not scalable to autonomous systems as they require a human actor in the loop to puppeteer avatar behavior. Recently, deep belief nets were utilized as a powerful yet flexible representation tool to model the variation and constraints of facial emotions and to produce convincing expression samples [27]. In [32] temporal restricted Boltzmann machines were used to transfer facial expression from one person to another target. While these approaches have shown promising results in transferring the same facial expression from one identity to another, they have not purported to model interaction dynamics of multiple person conversations.

In this paper, we studied human dyadic interactions to tackle the problem of facial expression generation in human-avatar dyadic interactions using conditional Gener-

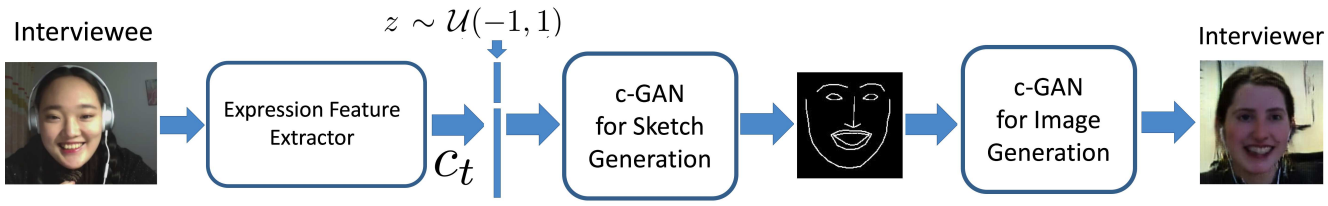


Figure 1. DyadGAN is composed of two stages of GANs, one to generate sketches and the other one to generate facial expressions of the interviewer. The inputs are facial expression features extracted from images of interviewees that serve as the conditioning vector c_t.

ative Adversarial Networks (GANs). Our goal is not simply to map the same facial movement of one individual to another but rather to build a model that takes into account behavior of one individual in generating a valid facial expression response in their virtual dyad partner. Previous work based on GANs and conditional GANs has shown success on producing synthetic images in various object categories and on predicting possible future frames in video sequences, including face images with different facial attributes such as emotion states, appearance cues or temporal information [24]. Our work differs from these in that we do not employ conditions related to facial attributes of generated identities, but consider the relationship and influence of one person's facial expressions in the reaction of the other. To this end, we introduce a two level optimization of GANs in interviewer-interviewee dyadic interactions, as shown in Figure 1. In the first stage, we generate expressive face sketches of the interviewer conditioned on facial expressions of the interviewee. The second stage generates complete face images conditioned on the face sketches generated during the first stage. This two stage approach allows us to learn an intermediate representation, the expressive face sketch, which could also be used to generate real facial expression images for a number of different identities. We demonstrate that our model is effective at generating visually compelling face images in dyadic interactions. Moreover we quantitatively show that the facial expressions depicted in the generated interviewer face images reflect valid emotional reactions to the interviewee behavior.

2. Related Work

Significance of facial expressions in dyadic interactions. Communication involves both verbal and nonverbal ways of making sure our message is heard. A simple smile can indicate our approval of a message, while a scowl might signal displeasure or disagreement. Moreover, the sight of a human face expressing fear elicits fearful responses in the observer, as indexed by increases in autonomic markers of arousal [26] and increased activity in the amygdala [15]. This process whereby an observer tends to unconsciously mimic the behaviour of the person being observed [2, 19] has been shown to impact a variety of interpersonal activities such as collaboration, interviews and negotiations among others [1, 2, 28, 11]. The classic study by Word et al. [30] demonstrated that interviewees fared worse when they mirrored less friendly body language of the interviewer, compared to what they did in the friendly condition. In parallel with the unconscious face processing route there is a conscious route, which is engaged, for example, when volunteers are explicitly asked to identify facial expressions or to consciously use facial expression as communicative signals in closed loop interactions. In many situations, an additional cue (an ostensive signal such as briefly raised eyebrows when making eye contact) is produced to indicate that the signaling is deliberate [7, 17].

Conditional generative adversarial networks . Conditional GANs [10, 23] are generative models that learn a mapping from random noise vector z to output image y conditioned on auxiliary information x: $G : \{x, z\} \rightarrow y$. A conditional GAN consists of a generator $G(x, z)$ and a discriminator $D(x, y)$ that compete in a two-player minimax game: the discriminator tries to distinguish real training data from generated images, and the generator tries to fail the discriminator. That is, D and G play the following game on $V(D, G)$:

$$\min_G \max_D \mathcal{V}(D,G) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)] + \mathbb{E}_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1 - D(x, G(x,z)))]. \quad (1)$$

In [24], various techniques from Deep Convolutional Networks have been introduced into GAN models to fix some of the previous limitations and produced compelling results. Previous works has employed different auxiliary conditional information such as labels [23], text [25] and images [14]; GANs based methods have tackled text-to-image/image-to-image translation [25, 14], face image synthesis [24], future frame/state prediction [21, 33], image manipulation guided by user constraints [34], style transfer [20] and 3D shape modeling [31].

3. Dyadic Dataset and Facial Expression Descriptors

As mentioned earlier, in this study we explored interviewer-interviewee dyadic interactions. Our dataset

consists of 31 interviews for undergraduate university admissions process. The purpose of interviews was to assess English speaking ability of the candidates. All participants were consenting adults that agreed to release of data for scientific research. The interviewees were prospective college students from a variety of ethnic backgrounds with a nearly even gender split (16 male and 15 female candidates). Each candidate was interviewed by the same interviewer (Caucasian female) who followed a predetermined set of academic and nonacademic questions designed to encourage open conversation and gather evidence of the candidate's English speaking ability. The interviews were conducted using Skype videoconferencing so the participants could see and hear each other and the video data from each dyadic interaction was captured. The duration of interviews varied from 8 to 37 minutes and the total dataset consists of 24 hours of video data (when including both interviewee and interviewer videos). It should be noted that since the interviewer is the same in each dyad we believe an advantage of this dataset is that it provides a significant amount of data under varying stimuli (31 different candidates) to adequately model the interviewer's behavior in this context.

Facial expression descriptor. We used Emotient's Facet SDK [12] to process the dataset and generate per-frame, 8-dimensional facial expression descriptor vectors, representing likelihoods of the following classes: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness and Neutral. Each value in the original descriptor represents the likelihood, in logarithmic (base 10) scale, of a target expression/affect being present. The evidence values typically range from -2 to 2 and can be easily transformed into the probability of an expression/attitude as follows:

$$P = \frac{1}{1 + 10^{-r}}. \tag{2}$$

In table 1, we illustrate the computed expression probability values of two participants in a dyad in Figure 3 at Time 00:17. We used the transformed facial expression descriptor d containing calculated probabilities of emotions.

4. Approach

Humans display a wide range of facial expressions, often with subtle differences to convey highly contextualized emotional states. We address the challenge of generating this highly complex process with a two stage model. The first stage is a conditional deep convolutional generative adversarial network (DC-GAN) [13] designed to produce expressive facial sketch images of the interviewer that are conditioned on the interviewee's facial expressions. The second stage is another DC-GAN to transfer refined sketch images into real facial expression images. In addition to being a more computationally tractable optimization problem, our two stage approach has the added advantage of learning an intermediate representation, the expressive facial sketch, which could be used to generate real facial expression images for a number of different identities.

On both stages we adapted generator and discriminator architectures from [24] and used modules of the form convolution-BatchNorm-ReLu [13] to stabilize optimization. In the training phase, we followed the standard approach to use mini-batch SGD and apply the Adam solver. To avoid the fast convergence of discriminators, generators were updated twice for each discriminator update, which differs from original setting [24] in that the discriminator and generator update alternately.

4.1. Expressive face sketch generation

Network architecture. Figure 2 summarizes the training procedure of the first stage. In the generator G, at first a 100 dimension noise z is sampled from the uniform prior $\mathcal{U}(-1, 1)$ and encoded with a temporal facial expression feature c_t computed from interviewee videos as shown in equation (3). For each frame of the interviewer at time t we considered the facial expression descriptors of the associated interviewee data between $[t - \delta t, t]$. We empirically used a time weighted average on all expression descriptors in $[t - \delta t, t]$:

$$c_t = \frac{\sum_{\tau \in [t-\delta t, t]} w_\tau d_\tau}{\sum_{\tau \in [t-\delta t, t]} w_\tau}, \tag{3}$$

$$w_\tau = \exp(\frac{\tau - t}{\delta t}). \tag{4}$$

Each element in c_t was normalized to $[0, 1]$ before we used c_t as input conditional vectors in the first level of our model.

The input is passed to two fully connected layers followed by batch normalization and rectified linear (ReLU) processing. The inference then proceeds as in a normal up-sampling layer followed by a Sigmoid function. In our model, the auxiliary information c_t is combined with intermediate features in all layers to magnify its influence: in full connection layers, c_t is simply concatenated with input/output features; in up-sampling layers, c_t is replicated spatially and depth-concatenated with feature maps.

In the discriminator D, at first a real or fake (generated) sketch image is depth concatenated with c_t. The combined input goes through two layers of stride-2 convolution with spatial batch normalization followed by leaky ReLU. Again two full connection layers are employed and the output is produced by a Sigmoid function. Similarly, the facial expression feature is concatenated with features in all layers in the discriminator.

Sketch training set generation. On each image frame of the interviewer sampled from dyadic interview videos, a face shape predictor similar to Kazemi et al. [18] is utilized

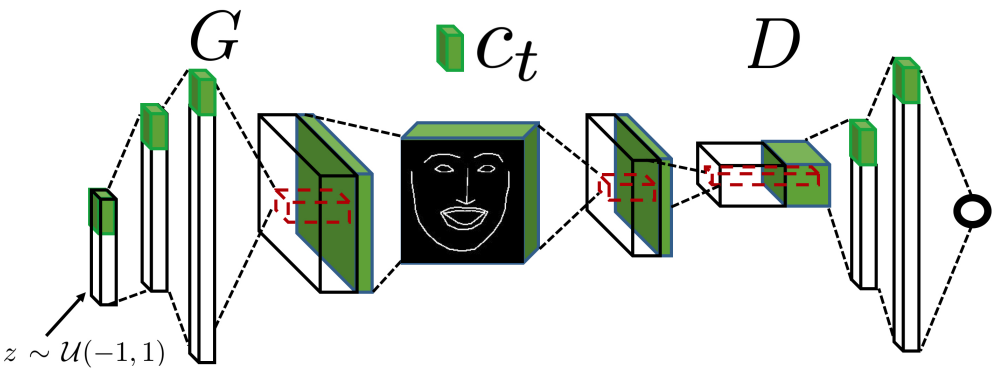


Figure 2. The architecture of the first stage, showing how the expression feature is concatenated in each feature layer of the generator and discriminator.

to obtain a set of 68 landmark points which depict components of human faces, such as eyes, eyebrows, nose, mouth and face out-contour. These landmarks were then linked by piece-wise linear lines of one pixel width. Figure 1 demonstrates the procedure of generating a sketch image from a sampled frame of the interviewer.

(a) (b) (c)

Figure 3. The interviewer and the interviewee interacted through online video-conferencing. Cropped image pair (a), (b) and (c) shows the corresponding facial expressions at 00:17, 07:10 and 11:33 respectively.

4.2. Sketch to image generation

Network architecture. We adopted the framework of Isola et al. [14] to learn the transformation of generated face sketches to complete facial expression images. To fulfill the training of this stage and construct sketch-image pairs, we used the sketch training set for the first stage as input to the generator and corresponding sampled facial expression images as the input to the discriminator. For each pair the sketch is strictly aligned with the corresponding face image, A sketch is passed through an encoder network containing 8 down-sampling layers and then a decoding network composed of 8 up-sampling layers to produce an image. To share the low-level information between input and output, a U-Net strategy [14] is utilized to concatenate corresponding feature maps of encoding and decoding layers, that is, all feature maps at the layer i of the encoding network are combined with those at layer ni in the decoding networks, where $n = 8$.

In the training phase of our sketch to image generation stage the GAN objective in equation 1 is combined with a L_1 loss to enhance image quality of outputs:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[||y - G(x, z)||_1], \quad (5)$$

$$G^* = \mathcal{L}(D, G) + \lambda\mathcal{L}_{L1}(G), \quad (6)$$

where G^* is our final objective. In practice we fixed $\lambda = 100$ and found it worked well. In such an approach, the discriminators task remains unchanged, i.e., distinguish real facial expression images from generated ones, but the generator's job is to not only fail the discriminator, but also to produce images matching the real samples y (the input to the discriminator) in an L_1 sense. The noise signal of z is not explicitly fed into this stage; instead randomness is only provided in the form of dropout, applied on first 3 layers in the encoding network of the generator at both training and inference time.

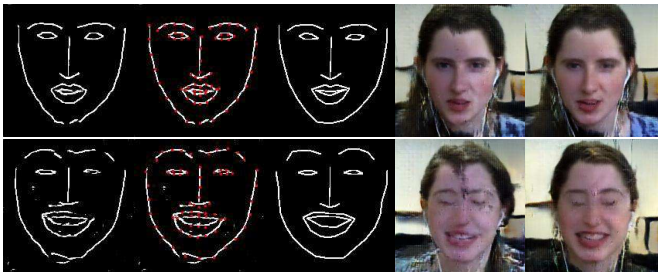
Input sketch denoising by landmark relocating. As introduced above, the sketches used in training are produced by linking the landmarks detected from real images. The lines in these sketches are noise-free and strictly with a width of one pixel. However, during the inference phase noisy sketches generated from our first stage (expressive face sketch generation) are used as inputs for reference. As shown in Figure 4, these noisy sketches deteriorate the quality of output images. To fix this issue, we trained a shape predictor by using pairs of sketches and landmark sets. We then deployed this shape predictor on noisy sketches to locate key landmarks and connected them with piece-wise linear lines to obtain cleaned sketches. Figure 4 demonstrates how our sketch de-noising process benefits the facial expression generation.

5. Experiments

Evaluating generative models that can sample but not estimate likelihood directly is a challenging problem.

	Joy	Anger	Surprise	Fear	Contempt	Disgust	Sadness	Neutral
Interviewer	**1.0**	2.45e-5	8.31e-7	8.07e-4	9.17e-10	9.41e-4	4.13e-9	8.50e-12
Interviewee	**1.0**	7.95e-6	6.81e-6	0.0166	4.05e-5	4.02e-4	3.82e-5	3.68e-8

Table 1. Expression/attitude probability scores of the interviewer and the interviewee in Figure 3 at Time 00:17 (first image pair). High level of Joy expression is consistent with the image content.



(a) (b) (c) (d) (e)

Figure 4. The de-noising process for generated sketches from Stage 1. Column (a) to (e) show the noisy sketches, the landmark relocation (denoted in red dots), the cleaned sketches, output facial expressions using (a) as input and output facial expressions using (c) as input. In the first row the initial sketch only contains very limited noise and broken lines on chin and mouth, but the cleaned sketch still largely suppress the defects in (d). In the second row the twisted and broken initial sketch fails to generate a reasonable face while the denoised sketch synthesizes an adequate one.

Goodfellow et al.[10, 23] used Parzen window-based log-likelihood to estimate probability of the test set data under p_g (the generators distribution). Radford et al.[24] applied its GAN model as a feature extractor on supervised datasets and evaluate the performance of linear models (such as SVM) fitted on top of these features. Those metrics do not assess the joint probability of x and y, therefore do not measure the effectiveness of our model to generate facial expressions in dyadic interaction. In order to quantitatively evaluate the quality of generated images, we organized two experiments.

In the first experiment we randomly sampled video clips of 5 seconds from interviewees and calculated their facial expression features c_t according to Equation 3. We input these feature vectors as conditions to our framework and generated facial expression frames of the interviewer. Once more we used Emotient's Facet SDK to extract two sets S_g and S_r. S_g contains facial expression descriptors d_t^g computed from the generated interviewer images and S_r contains descriptors d_t^r computed from the real interviewer images that temporally aligned with interviewee video clips. We then analyzed the statistical properties of these two sets to determine if they have significant differences.

In the second experiment we input canonical expression descriptors of interviewees (pure Joy, Anger, Surprise etc.) to analyze the generated responses of the interviewer. Our intuition is that a set of interviewer's facial expressions generated in reaction to a canonical emotion of the interviewee (e.g. Joy) would be similar to each other and different from those generated in response to a different canonical expression (e.g. surprise) of the interviewee.

	Population 1	Population 2
Number of Observation	8000	8000
Sample Mean	1.3049	1.3802
Sample STD	0.17875	0.067433
Significance level	0.05	
p-level	1.706×10^{-30}	

Table 2. Results of a lower-tailed, two sample t-test on Population 1 and Population 2.

5.1. Experiment 1

We randomly sampled 1000 short video clips of interviewees for each of eight major emotions in Table 1 and computed expression features c_t according to Equation 3. We enforced to only select those video clips whose last frame produces at least one emotion (in corresponding d) with a probability above 70%. In this way we obtained 8000 expression feature vectors c_t. For the real frames of interviewers that temporally aligned with those 8000 clips, we extracted expression descriptor d_t^r to form an set S_r.

By using each c_t, we randomly generated 15 images of interviewers, computed 15 face expression descriptors ($d_t^g[1]$ to $d_t^g[15]$). We calculated the Euclidean distance between d_t^r and d_t^g for 15 times and get the average distance dis_t. In this way we produced a population set (Population 1) of 8000 distance values which measure the expression difference between real interviewer images and generated interviewer images. Each distance value in this set corresponds to a specific c_t described above.

We also built another population set of distances (Population 2) to compare with Population 1. For each d_t^r, we randomly sampled 100 d^g which is not corresponding to c_t (of d_t^r). Again we calculated the Euclidean distance between d_t^r and those d^gs for 100 times and get the average distance dis_t^{rand}; in this way our Population 2 also contains 8000 distance values. The sample means and sample standard deviations of two populations are illustrated in Table 2. For simplicity, we ignored the correlations among those values and assumed the independence of them.

Our hypothesis is that the expression distance between a real interviewer image and generated images w.r.t. the same c_t should be smaller than that of randomly computed

	Joy	Anger	Surprise	Fear	Contempt	Disgust	Sadness	Neutral
Joy	**0.5770**	1.2014	1.1011	1.1787	1.1602	0.9533	0.9280	1.0610
Anger	1.2014	**0.9090**	0.9815	0.9230	1.0314	0.9877	0.9522	0.9166
Surprise	1.1011	0.9815	**0.9087**	1.0422	1.0571	1.004	0.9590	0.9820
Fear	1.1787	0.9230	1.0422	**0.8172**	1.037	0.9906	0.9962	0.9006
Contempt	1.1602	1.0314	1.0571	1.0370	**0.8169**	1.0873	1.0505	0.9054
Disgust	0.9533	0.9877	1.0038	0.9906	1.0873	**0.9222**	0.9020	0.9636
Sadness	0.9280	0.9522	0.9590	0.9962	1.0505	0.902	**0.8434**	0.8856
Neutral	1.0610	0.9166	0.9820	0.9006	0.9054	0.9636	0.8856	**0.7472**

Table 3. Comparison on intra-set and inter-set average distances between different sets of the second experiment.

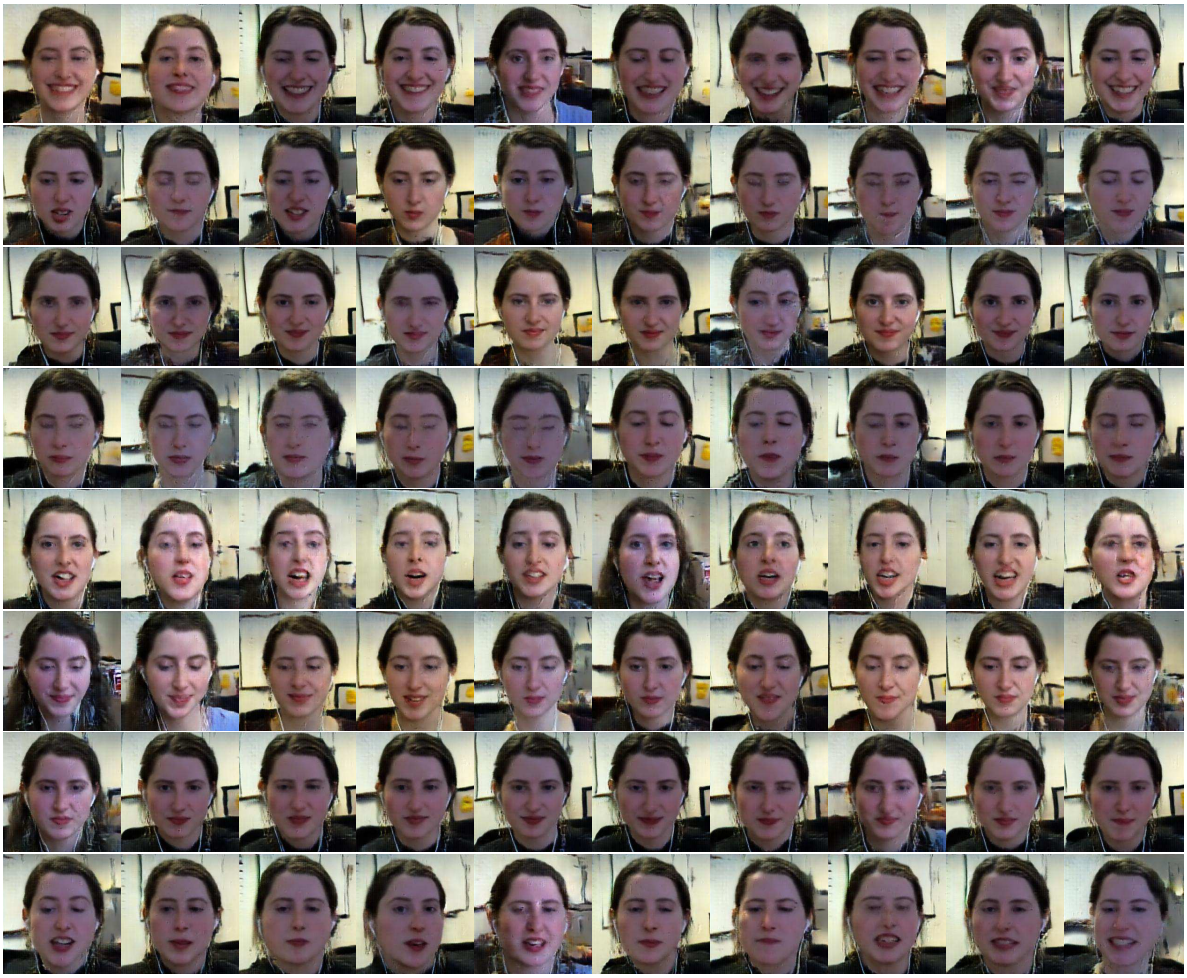


Figure 5. Each row represents exemplar generated expressions of the interviewer when interacting with eight canonical emotions of interviewees: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sad and Neutral. These samples demonstrate that the generated interviewer response is qualitatively consistent for each of the interviewee emotions.

one in the control set (Population 2). This could be tested by using a lower-tailed, two sample t-test in which the null/alternative hypotheses is defined as H_0: $\mu_1 = \mu_2$ and H_α: $\mu_1 < \mu_2$ respectively, in which μ_1 (μ_2) represents the mean of Population 1 (Population 2). We adopted Matlab function *ttest2* to conduct this test in which Satterthwaite's approximation [29] was used for the case that equal variances of two data populations are not assumed. As shown in Table 2, at a significance level of 0.05, the computed p-value is very close to 0. So we accept the alternative hypothesis H_α ($\mu_1 < \mu_2$) at the 0.05 significance level, which concludes a statistically significant reacting effect of our generative model.

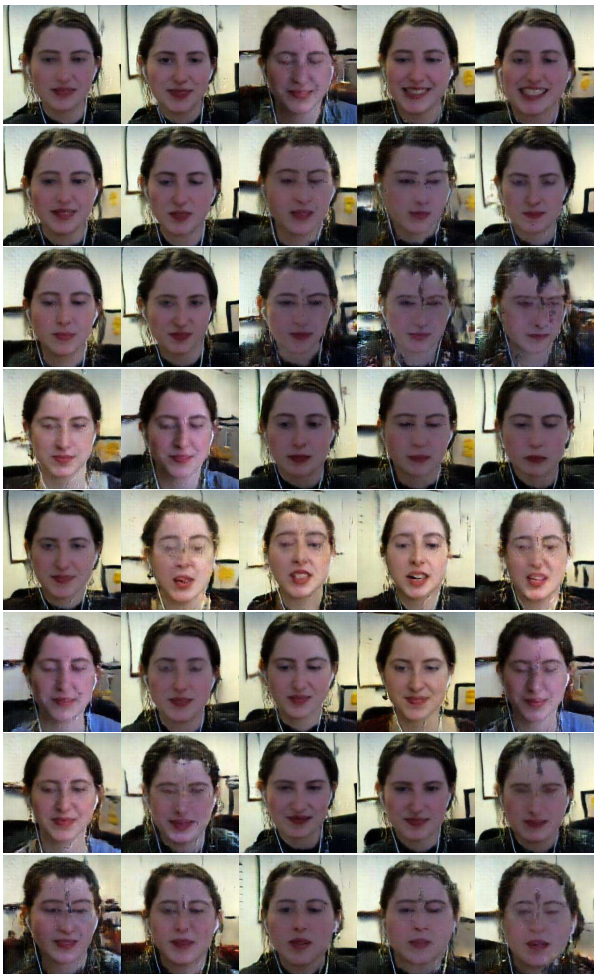


Figure 6. Intensity of Expression: Each row represents exemplar generated expressions of the interviewer when interacting with interviewee emotions of varying intensity. Top to bottom: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sad and Neutral. Left to right, the emotion intensity: 0, 0.25, 0.50, 0.75 and 1 respectively.

5.2. Experiment 2

In this experiment we test how consistently our model generates valid interviewer facial expression images in response to interviewee behavior. A set of 8 canonical interviewee facial expression features c_t are fed to our model as one-hot vectors, for instance $[1, 0, 0, 0, 0, 0, 0, 0]$ representing interviewee Joy expression with a probability of 1, $[0, 1, 0, 0, 0, 0, 0, 0]$ representing Anger and so forth. We then use our GAN model to randomly generate 1000 interviewer expression images for each interviewee canonical expression feature. Figure 5 shows some examples from these generated images. It can be observed that for each interviewee canonical expression, the generated interviewer response (images on the same row) is valid and consistent with each other. To verify this quantitatively we extract facial expression descriptors d^g from the generated interviewer face images (see section 3 for facial expression descriptor details) and group these into 8 sets corresponding to each of the canonical interviewee facial expressions. Table 3 shows the average Euclidean distances of the facial expression descriptors from one set to another. As can be seen the intra-set average distance on the diagonal of Table 3 (computed between descriptors of the same set i.e. interviewer response to one of the 8 canonical interviewee expressions) is generally smaller than inter-set average distances.

Facial expression transition effect. To demonstrate the ability of our system to generate consistent facial expressions of the interviewer w.r.t interviewees' expression features of different intensity levels, for eight expression inputs we presented an expression transition effect in Figure 6. For Joy, a series of 5 interpolation expression features between $[0, 0, 0, 0, 0, 0, 0, 0]$ and $[1, 0, 0, 0, 0, 0, 0, 0]$ illustrate that the interviewer not only 'reacted' to interviewees' joy expressions according to the intensity level, but also exhibited a smooth transition from left to right.

6. Conclusion

In this paper we have presented an approach to generate face images depicting contextually valid facial expression during interviewer-interviewee dyadic interactions. A key novelty in our approach is that the face image generation is conditioned not on the generated identity's own/self attributes but rather the facial expressions of their dyadic conversation partner. We believe this allows us to better capture the influence of dyad partner's behavior in generating a response. It should be noted that since our model was trained on a dataset consisting of a single individual interviewing many candidates, inferences drawn from our experiments do not necessarily generalize to a multitude of interviewer personalities. To extend our approach to multi-interviewer scenarios, in addition to having a larger dataset with multiple interviewer identities, standard style transfer techniques could be utilized, or more sophisticated shape registration methods could be performed to align face shapes of different identities to a tangent space before the GAN training. To enhance the generation quality, different forms of loss function could also be used to better regularize the GAN objective. Finally, our approach can be combined with a temporal recurrent networks such as LSTM to synthesize continuous video frames of facial expressions in dyadic interactions.

7. Acknowledgment

We would like to thank Scott Clyde and Jon Burdick of University of Rochester for leading the interview data collection activity in collaboration with ETS and for sharing this data for research study.

References

[1] S. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47:644–675, 2002.

[2] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction*, Chicago, 2013.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. SIGGRAPH '99, 1999.

[4] T. D. Bui, D. Heylen, M. Poel, and A. Nijholt. Generation of facial expressions from emotion using a fuzzy rule based system. In M. Stumptner, D. Corbett, and M. Brooks, editors, *AI 2001: Advances in Artificial Intelligence*, pages 369–391. Springer, Berlin, 2001.

[5] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. SIGGRAPH '94, 1994.

[6] L. M. D. Devault, A. Rizzo. Simsensei: A virtual human interviewer for healthcare decision support. In *Thirteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2014.

[7] D. W. D. Sperber. Relevance: communication and cognition. *2nd edn.Oxford, UK: Blackwell*, 1995.

[8] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[9] C. Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3453-3458*, 2009.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2014.

[11] Y. Huang and S. M. Khan. Mirroring facial expressions: Evidence from visual analysis of dyadic interactions. In *ICMR '16 Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*.

[12] iMotions Inc. Manual of Emotient's Facet SDK. 2013.

[13] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv e-prints*, Feb. 2015.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv e-prints*, Nov. 2016.

[15] R. J. D. J. S. Morris, A. Ohman. A subcortical pathway to the right amygdala mediating unseen fear. In *National Academy of Science*, 1996.

[16] M. S. C. T. . M. N. J. Thies, M. Zollhofer. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[17] U. F. K. K. Kampe, C. Frith. hey john: signals conveying communicative intention toward the self activate brain regions associated with mentalizing, regardless of modality. *Neuroscience*, 2003.

[18] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[19] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

[20] C. Li and M. Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. *ECCV*, Apr. 2016.

[21] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ArXiv e-prints*, Nov. 2015.

[22] M. M. J. M. R. P. M.E. Hoque, M. Courgeon. Mach: My automated conversation coach. In *15th International Conference on Ubiquitous Computing (Ubicomp)*, 2013.

[23] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *ArXiv e-prints*, Nov. 2014.

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv e-prints*, Nov. 2015.

[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. *ArXiv e-prints*, May 2016.

[26] A. O. . J. J. Soares. Emotional conditioning to masked stimuli: expectancies for aversive outcomes following nonrecognized fear-relevant stimuli. *Experimental Psychology*, 1998.

[27] J. M. Susskind, G. E. Hinton, M. J. R., and A. K. Anderson. Generating facial expressions with deep belief nets. In *In Affective Computing, Emotion Modelling, Synthesis and Recognition*. I-Tech Education and Publishing, 2008.

[28] A. A. Tawfik, L. Sanchez, and D. Saparova. The effects of case libraries in supporting collaborative problem-solving in an online learning environment. *Technology, Knowledge and Learning*, 19(3):337–358, 2014.

[29] B. L. Welch. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

[30] C. O. Word, M. P. Zanna, and J. Cooper. The nonverbal mediation of self-fulfilling prophecies in interracial interaction, 1974.

[31] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling.

[32] M. D. Zeiler, G. W. Taylor, L. Sigal, I. A. Matthews, and R. Fergus. Facial expression transfer with input-output temporal restricted boltzmann machines. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, pages 1629–1637, 2011.

[33] Y. Zhou and T. L. Berg. Learning Temporal Transformations From Time-Lapse Videos. *ECCV*, Aug. 2016.

[34] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.