

Measuring Emotion Intensity: Evaluating Resemblance in Neural Network Facial Animation Controllers and Facial Emotion Corpora

Sheldon Schiffer^a

Department of Computer Science, Occidental College, 1600 Campus Road, Los Angeles, U.S.A.

Keywords: Autonomous Facial Emotion, Emotion AI, Neural Networks, Animation Control, Video Corpora.

Abstract: Game developers must increasingly consider the degree to which animation emulates the realistic facial expressions found in cinema. Employing animators and actors to produce cinematic facial animation by mixing motion capture and hand-crafted animation is labour intensive and costly. Neural network controllers have shown promise toward autonomous animation that does not rely on pre-captured movement. Previous work in Computer Graphics and Affective Computing has shown the efficacy of deploying emotion AI in neural networks to animate the faces of autonomous agents. However, a method of evaluating resemblance of neural network behaviour in relation to a live-action human referent has yet to be developed. This paper proposes a combination of statistical methods to evaluate the behavioural resemblance of a neural network animation controller and the single-actor facial emotion corpora used to train it.

1 INTRODUCTION

As expensive as they are to design and produce, photo-realistic human agents have become a common attraction in contemporary video game design of non-player characters (NPCs). To get them to behave with emotional veracity, video game developers are using AI techniques to control facial expressions. Developers may choose between at least two approaches. The first evolved through the Computer Graphics research community. It prioritizes mimetic resemblance of movement and modelling to the appearance of a performing actor or model. The second was developed by the Affective Computing community, and prioritizes emotional resemblance, which is the ability of the avatar to autonomously elicit a facial expression based on an integrated emotion model. These two approaches have evolved over several decades using different models of simulation. The former creates a system of expression generation based on appearances on the surface of the agent's face. The latter attempts to encapsulate an emotion generation system that is located "inside" the agent. Both approaches rely on the same two components: (1) a collection of video samples of the face from which to extract structured data about a

subject's facial state and (2) a Neural Network (NN) controller trained from the aforementioned data, and programmed to control an avatar's facial mesh that resembles the face of the performing actor found in the collection of video samples.

In this paper we ask, how does a researcher evaluate the resemblant quality of a NN facial animation controller and the facial emotion video corpora on which it was based? An evaluation technique must determine if the video corpus and NN architecture that drives an NPC's facial expression animation behave in an objectively similar manner to the original actor's facial elicitation as depicted in the video corpus. Past research in the graphics community has focused evaluation procedures on the error reported by algorithms that render resulting animation frames in relation to pre-defined visemes or expressions. But autonomous emotion in a virtual agent cannot be performed precisely the same way for every stimulus. Such consistency would be perceived as uncanny and mechanical. Thus, a statistical approach that describes a range of acceptable "error" is what we propose as unknown probabilistic causes of variation in facial expressions.

A concern for accuracy is also shared by Affective Computing researchers. The process of validating the accuracy of facial emotion elicitation video corpora

^a <https://orcid.org/0000-0001-5862-5239>

has primarily been used for research in the production of NNs for Facial Emotion Recognition (FER) software systems. The primary intention of corpus validation has been to warrantee that the emotion label value assignments for each frame for static images, or each clip for dynamic images, is statistically consistent. The system of giving intensity values to emotion names identified in static photos evolved from a century-old method of recognition techniques (Darwin and Prodger, 1872/1998). Contemporary emotion recognition classifies facial muscle group behaviours into culturally and linguistically determined emotion names, or “labels” (Ekman 2006). The use of FERs provides a ground truth referent on which to model the facial expressions for NPCs. Developers of games and interactive media need a method to determine if the two components that influence the behaviours of an animated character or agent – the NN and the video corpus that trained it – are producing facial emotion elicitation as intended. Thus, video game developers of photo-realistic characters can draw from the techniques of both graphical and affective computation to determine the emotionally resemblant quality of their corpora and NN. Using some aspects of both approaches, a method of corpora production and evaluation can provide consistently evaluated data sources for training NN controllers. Two statistical techniques are proposed that provide a preliminary basis for analysis.

2 RELATED WORK

Research in computer graphics and affective computing were consulted to develop a process of evaluating resemblance derived from NN controllers and the corpora used to train them.

2.1 Example-Based Animation

New methods of simulating facial elicitation in Computer Graphics prioritize graphical accuracy of modelling and animation over emotional autonomy. Several studies by Paier et. al propose a “hybrid approach” that use “example-based” video clips for frame-by-frame facial geometry modelling, texture capture and mapping, and motion capture (Paier et al., 2021). In their experiment, a performing actor speaks a few lines or elicits a set of idiosyncratically defined gestures. The recording or real time live stream provides information for automatic geometry and dynamic facial texture generation (Paier et al., 2020). Then, a NN using a variable auto-encoder (VAE)

integrates motion for mesh deformation, while another NN selects animation sequences from an annotated database. Database annotation of animation has demonstrated the efficacy of movement data classification of a single actor that can be used later for semi-autonomous expressions utterances. Their approach demonstrates highly resemblant avatar animation for short single-word utterances or single-gesture elicitation.

An assumption that using speech as a primary modality for determining emotional states, belies the belief that facial elicitation is more reliably understood as a function of word utterance. The emphasis on speech synchronization assumes that the expressive meaning of an intended facial elicitation will correspond to the semantic context of the spoken word. This emphasis is found in a study by Suwajanakorn et al. (2017) that uses the vast collections of video samples of a U.S. President. From a 17-hour corpus, the investigators mapped speech from persons who were not the subject of their video corpus, onto a moving and speaking face of the presidential subject. Their method discovered that optimal training of their NN benefited from expression positions of the face of both past and future video frames to best predict how to synthesize deformations of the mouth right before, during and at the completion of spoken utterances. Thus, their NN incorporated Long Short-Term Memory (LSTM) cells to predict mouth animation synthesis for the video of upcoming visemes. For the experiments conducted for this research, we also deployed LSTM cells and found them useful for the same benefit.

From the standpoint of a designer of autonomous agents for video games however, neither approach mentioned thus far provides a model of fully autonomous elicitation in response to measurable stimuli. Both examples show that the use of single-actor or single-subject corpora is viable for training a NN to simulate the facial expressions of an actor’s character design or that of a real person. Viability is made possible with a NN that learns the dynamics of facial expression based on labelled visemes. This technique we also integrated through training with multiple video clips of an actor repeating a performance in reaction to the same stimuli. This approach proved useful in our development of NNs targeted to train specific emotions.

Another distinction in the Computer Graphics approach is that their corpora structures do not correlate with widely used psychological classifications of emotions and their elicitation (e.g. the six to twelve basic emotions identified by social and computational psychologists). Neither Paier’s nor

Suwajanakorn's research disclose a classification system of emotions, and therefore the meaningfulness of their synthesized expression rely on arbitrarily selected spoken semantics rather than independently systematized semantics of facial expression. The graphics approach instead prioritizes frame-by-frame facsimile of labelled visemes on a real human source as displayed by its simulated avatar. Nonetheless, the mimetic quality of results created by the workflows of both Paier et al. and Suwajanakorn et al. must be considered for autonomous facial emotion elicitation.

2.2 Performance and Appraisal Theory

Affective Computing has been much more focused on the history of encoding models of emotions. Automatic facial expression generation evolved over several decades, lead in part by computational psychologists and developers. Academic software developers of virtual agents designed synthetic emotion elicitation systems for fundamental posture, gesture and facial expressions that infer emotional states. The Oz Project, a collection of video game experiments and research papers realized by Loyall, Bates and Reilly in the late 1990s, made use of emotion generation processes actors use to prepare and train for performance (Loyall, 1997). The Method approach, a series of practiced exercises as developed by Russian theater director Constantin Stanislavsky (Moore, 1984), were combined in the Oz Project with the emotion system structure of Ortony, Clore & Collins (OCC) known as Appraisal Theory (Ortony et al., 1990). Together, these approaches were implemented in the virtual agents of their experiments (Bates et al., 1994). While the Oz Project made use of a theory for motivating emotion elicitation for AI agents, it did not use an exhaustive dataset from which to draw machine learnable examples into a model, as recently evolved by Schiffer et al. (2022).

2.3 Facial Corpora Production

The primary use of facial emotion corpora is to provide a ground truth baseline for general emotion recognition systems. This goal is different than creating a corpus for simulation systems. Therefore, our research had to adapt the progress of recognition applications for our intended simulation application.

Published corpora reports indicate their baseline definitions of static and/or dynamic emotion elicitations of the human face. Distinctions between corpora consist of two fundamental feature categories: the method of production of the video clips, and the method of validation of the corpus. Clip

production or selection methods diverge in the choice to use actors as practiced by Bänziger et al. (2011) versus non-actors (Vidal et al. 2020). Busso et al. (2017) produced a corpus with tightly scripted scenarios, while Metallinou et al. (2010) used more improvisational techniques. As our ultimate goal was to systematize simulated facial elicitation, we sought to discover how actors with a scripted scenario written to generate emotions were used in facial emotion corpora. We considered how Lucey et al. (2010) and Bänziger et al. (2011) produced clips in a controlled studio environment to support an actor's undistracted preparation techniques.

Since our objective was simulation of a limited and targeted set of facial elicitations, evaluation methods we developed required fewer elicitation variations than a corpus or NN designed for generic emotion recognition. Nearly all corpora reference the Facial Action Coding System (FACS) that correlates groups of muscles, called Action Units (AUs), to manipulations of the face to form expressions of at least six basic emotion labels: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* (Cohn et al. 2007). These labels are used to estimate emotion intensity or provide perceived levels of arousal and valence (Soleymani, 2014). Our approach used the classification capacity of FERs to develop single-actor facial emotion corpora on a targeted set of emotions.

Drawing from the work of Bänziger et al. (2011), our system similarly segregates all emotion label values by intensity on the Russel Circumplex Model (Posner et al., 2005). We sought to train NNs for specific emotion labels where each would be designated to control a set of facial AUs of an avatar's wireframe mesh modeled after the face of a performing actor of the corpus. This approach allows for classifying resemblance by evaluating the intensity difference between the emotion label values of the performing actor and their animated avatar.

2.4 Neural Networks for NPCs

Kozasa et al. (2006) showed a preliminary use of an affective model for an emotive facial system in an NPC based on a dataset of expressions. Theirs used a 3-layer feed-forward artificial neural network to train an NPC from "invented" data for parameters fed to a NN model as they claimed no databases at the time existed to train their model. Later, using appraisal theory-based design from virtual agents, Mascarenhas et al. (2021) integrated the FATiMA architecture with a NN model in educational games. Khorrani et al. (2016) show that the use of LSTM cells for emotion recognition of facial video was

shown to improve previous NN performance for emotion recognition. The method proposed in this paper also adopts Long Short-Term Memory (LSTM) cells in a NN. Unlike the methods that this paper’s research proposes, these previous works did not use NN models whose emotion elicitation training is drawn from single-actor video corpora, but instead chose corpora with clips of multiple human subjects.

3 PRODUCTION METHODS

Unlike many corpora, this research uses a single actor as the subject of corpus. The intended use is to train a NN to control facial animation of a photo-realistic NPC in a 3d video game. The NPC becomes the actor’s character Avatar. The general usefulness of this research is the method of corpora production and evaluation. A brief overview of our corpus production and neural network design follows.

3.1 Corpus Production

There are two phases for our corpus production: first, designing a dyadic conversational scenario and second, rehearsing and recording video clips. Scenario design consisted of two characters for actors to perform asynchronously following a *dialog behaviour tree* in the form of a directed acyclic graph. Actors were cast and rehearsed in preparation of the video clip recording. One actor played the Stimulus Source character and recorded a video edited beforehand of all path variations as if they were addressing the other character, the Emotion Model. Then the Emotion Model performed back to the camera reacting to synchronized video from the pre-recorded performance paths of the Stimulus Source as shown in Figure 1.

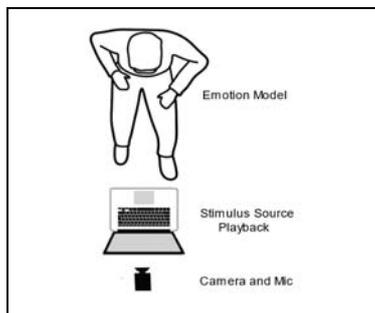


Figure 1: Setup for sample recording.

The design of the dialog behaviour tree consisted of distinct fixed start and end nodes with three layers

of six nodes in between. These intermediate layers allow two possible nodes of dialog turns. No node or edge could be repeated within a path of the tree as illustrated in Figure 2. The rules of the graph allow 32 paths through the tree. Each edge segment (the circle labelled letters) of the tree had a targeted emotion label. Segments used the same lexical content from the tree, though all paths had distinct dialog sequences. With the variations in paths, the *dialog behaviour tree* created a permutable performance structure with stimuli for elicitations to occur.

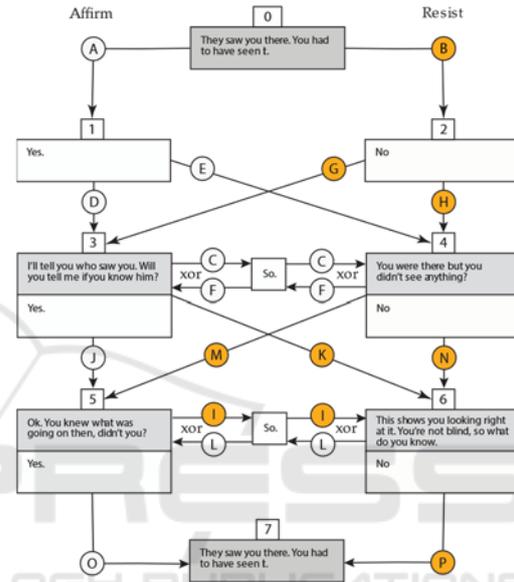


Figure 2: Box nodes at dialog turns (3, 4, 5, 6) and monolog events (0, 1, 2, 7). Edges represent mental actions. Orange-colored nodes were used in the data analysis. As an acyclic graph, a path can use edge C xor F and I xor L.

Each path was recorded 9 times using three distinct degrees of intensity of action: high, medium, and low, to provide more variation to train the NN as practiced by Wingenbach, Ashwin and Brosnan (2016). Thus 9 clips times 32 paths yielded 288 total clips of the Emotion Model.

3.2 Post-Processing Emotion Analysis

Each clip was post-processed by the FER, *Noldus FaceReader 8*, for frame-based emotion analysis. The output data consisted of 7 normalized emotion label values for each of six emotions, *happiness, sadness, anger, fear, surprise, disgust* plus *neutral*. A new tuple of seven normalized values were output 3 times each second. Noldus FaceReader is a tested and ranked FER system that has produced emotion recognition validation results that match the accuracy of human annotators (Skiendziel, Rösch and

Schultheiss 2019). Furthermore, the recognition accuracy rate of Noldus FaceReader has been documented as high at 94%. Its output is machine-readable text consisting of tuple instances of emotion recognition scores for each frame of video. These scores became the data used for training the NN using the Python language and the TensorFlow library. The FER was used to analyse all emotion values, plus neutral.

3.3 From Emotion Model to Avatar

The Emotion Model was orthographically photographed from overlapping angles to produce a photorealistic head mesh that resembles the actor. The head mesh was generated by the *FaceBuilder* plugin for the 3d animation software system *Blender*. *FaceBuilder* is a modelling tool for supporting 3d head animation with a facial rig whose vertex groups are controlled by shape key actuators within Blender. These shape keys were designed to move the same alignments of facial muscle groups defined in the AUs of FACS. The head mesh and the shape keys embedded in the facial rig were deployed in the game engine Unity 2022. The shape keys were put into autonomous motion by programmable blend shapes in Unity that receive streamed emotion data from the NN animation controller responding to the face of the Stimulus Source. The embedded NN receives the FER data and “reacts” to it in a way that intends to statistically resemble the character behaviour the actor created in the video clips in the single-actor corpus. The NN-generates prediction data in the form of normalized emotion label values as a means of autonomously controlling the *FaceBuilder* head mesh to animate facial expressions.



Figure 4: Developing Emotion Model avatar with Blender plugin Keen Tools *FaceBuilder*.

3.4 Recurrent Neural Network Design

Among the clips generated for this research, 68.8% of the corpus (198 clips) was used only for training the NN. 20.1% of the corpus (58 clips) was used only for

validation. The remaining 11.1% of the corpus (32 clips) was used to test the NN’s behavioural resemblance to the actor corpora on which it was trained.

The principal components of the NN follows a Recurrent Neural Network (RNN) design. Each component of the neural network was selected for its probabilistic ability to choose values of coefficient weights and biases for specific input features of the data that the NN was trying to predict. Predicting the facial elicitation of game characters based on training data from an actor’s performance requires spatial and temporal data representation. For our experiments, facial feature positions were estimated from their spatial contexts using Dense architectures (fully connected). We used a Dense layer of perceptrons that were fed two layers of bi-directional LSTM cells. The LSTM layers auto-regressively receive data from 10 seconds in the past using 3 instances of emotion label data per second. Since the data for this experiment was fed pre-processed emotion data tables (as opposed to a live video stream), the NN analysed 10 seconds into the future as well. Temporal relations of elicitation events in the data were processed by LSTM cell layers, while spatial relations of facial features were handled by the Dense cell layer. Each emotion label was assigned its own NN, so the designed recurrent NN was cloned into a team of 7 NNs and trained on synchronized data generated from each elicited emotion from the Emotion Model and the Stimulus Source.

4 EVALUATION METHODS

The evaluation methods proposed provide the developer of NPCs a quantitative process that measures behavioural difference. Optimally minimized difference in data can be interpreted as statistical *resemblance*. It is our intention to demonstrate statistically, that given the same or similar stimulus, prediction data from the NN can control animation that resembles observed FER-generated data of the human Emotion Model on which it was trained. The resemblance then depends on minimizing the amount of error between predicted behaviour performed by the Avatar and observed data performed by the Emotion Model. But the predicted data is not a single set for each instance in time. Instead, each time-instance within a path through the dialog behaviour tree is a video frame shared by at least 9 video clips and their edge segments, as well as other paths that share the same edge segment. Therefore, since all the video clips are precisely

synchronized, each of the frames in the experiment has a mean emotion value drawn from at least 9 clips of the same edge segment. And this value can be used to calculate error in relation to the predicted value at that frame demonstrated by the NN. By looking at the difference within ± 1 standard deviation, two useful statistical properties provide the results to determine statistical resemblance. By calculating the Percentage of Extreme Residuals (PER) that fall outside of ± 1 standard deviation, a first test of resemblance can be applied to the prediction data. To support those results, the Root Mean Square Error (RMSE) provides an amplification of variance. If the *neutral* emotion label values are used as a benchmark, RMSE becomes an additional statistical property to show if a NN facial animation controller resembles the character facial movement the actor generated, and if the corpus that trained the NN is sufficiently robust to confirm resemblance to any emotion labels.

4.1 Percent of Extreme Residuals

For each emotion and for any segment or combination of segments of a path through the dialog behaviour tree, it is useful to know how many frames have mean emotion values that fall outside ± 1 standard deviation from the mean of observed values at that frame. For this research, the Percentage of Extreme Residuals (PER) is calculated as follows:

$$PER = \frac{\sum_{i=1}^n 1_{|p_i| > \sigma_i}}{n} : \pm \sigma_i = \sqrt{\frac{\sum_{j=1}^m (x_i - \mu)^2}{m}}$$

Where n is the size of the set of all predicted emotion values and p_i is the predicted value of each frame measured for emotion values in the dialog behaviour tree edge segment set. For each absolute value of p_i that exceeds σ_i of all emotion values at the i -th frame, increment the sum by 1 and divide the by n such that we define σ_i as the standard deviation at the i -th frame of an edge segment. To calculate $\pm \sigma_i$, the set m is a count of all video clips j that cover an edge segment at the i -th frame where x_i is the observed value at the i -th frame and μ is the mean of all values at the i -th frame of a given emotion for a given edge segment. PER is the first measurement to consider.

Validation of a NN facial animation controller should not require the facial mesh behaviour to exactly animate the same way every time when it receives the same input from the Stimulus Source. But how wide should a range of variance be to seem human-like? Consider ± 1 standard deviation. With the mean of any emotion value at each frame as a reference, a range of *resemblance* can be defined

around the mean by using the standard deviation. The experiments of this research found as *resemblant* the frames where the predicted emotion value of a given frame fall within ± 1 standard deviation from the mean. For this condition to occur, the predicted value will fall into a value space with at least 68% of the observed values. For frames that fall outside ± 1 standard deviation, they shall be determined as *not resemblant*.

4.2 Root Mean Square Error

Statistical methods proposed in this research consider the mathematical characteristics of non-linear regressive models deployed in NN design. The NN deployed in this research uses RNN components: a layer of LSTM cells that include *sigmoid* and *tanh* as component functions, both of which are nonlinear. Since the regression model embedded in the NN used non-linear functions to autonomously elicit emotion, the experiment used statistical methods that interpret variance and are suited to non-linear regression.

The Root Mean Square Error (RMSE) is used to measure facial elicitation resemblance as a function of variance between two synchronized time series data sets: observed test data and generated prediction data. Each of the differences between the observed and predicted values, referred to as residuals, aggregates their magnitudes from point to point in a data set. The resulting value is always positive where 0 is a lower bound representing a perfect fit between the observed and predicted data. The mean of observed values for each emotion at each frame was used as a baseline to compare amplified variance between the emotion labels. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - y_p)^2}{n}}$$

y_i is the observed value at the i -th frame. y_p is the predicted value for the i -th frame. n is the population count, or total number of values in the population of observed values. Under the radical sign, RMSE aggregates and amplifies the variance by squaring the summed difference before dividing by the population count n . A subset of squared differences will be magnified quadratically and summed once before being squared. Meanwhile, a subset of squared and summed smaller differences will diminish quadratically. Thus, the extreme differences, large or small, will be amplified. Since emotion values of video frames of facial emotion elicitation are normalized to fractions in a 0 to 1 scale,

magnification of variance provides a visible contour of the behaviour of residual values in relation to the mean of observed values. A low RMSE will lean toward greater *resemblance* provided the PER is also *resemblant*.

5 RESULTS

Statistical error scores were computed to determine the behavioural resemblance between the mean of emotion values at each frame for all of 6 emotions. See Table 1.

Table 1: Sums of Emotion Values.

Emotions Analysed on 126 Frames 14 Clips with Edges B, G, H, I, K, M, N, P		
Emotion	Frame Value Sum	% Neutral
Neutral	871.41	--
Angry	60.64	6.95
Disgusted	7.95	0.91
Happy	40.90	4.69
Sad	705.88	81.00
Scared	54.65	6.3
Surprised	54.40	6.24

Neutral was also computed and is used as a benchmark from which to evaluate the accuracy of other emotions. *Neutral* is the absence of emotion and is theoretically at 1.0 when all other emotion values are 0.0. *Neutral* nearly always has the highest summation of accumulated emotion values over time as its values increase each time the face returns to *neutral*-dominant positions during transition to and during the listening phase of dyadic conversation. In all tests for this research, neutral value summations exceeded all other emotion value summations for any edge segment. *Neutral* therefore has the highest probability of yielding the highest value instances of any randomly analysed frame.

The highest summation of observed values should provide the lowest percentage of PER errors for the predicted values of emotion elicitations. As shown in Table 2, *neutral* PER is 0.0794, the lowest of all recognizable labels. Therefore, following the behaviour of *neutral*, the next highest PER emotion may provide proportionally *resemblant* results, proportional in that the higher the percentage the emotion's sum is to the sum of neutral, the more *resemblant* the primary emotion values are in relation to the *neutral* scores. Table 1 shows the emotion with the highest percentage close to neutral for frame value sums is *sadness* with 81%. The next highest is *anger* at 6.95%. With such a distant second position and so

far off the benchmark of *neutral*, one should doubt the resemblance of *anger*, while taking note of *sadness*.

Table 2: Error for Edges B, G, H, I, K, M, N, P.

Emotions Analysed on 126 Frames Error Between Mean of Observed and Predicted Data				
Emotion	meanSD	RMSE	meanRMSE	PER
Neutral	0.1680	0.1857	0.1791	0.0794
Anger	0.0516	0.0783	0.0705	0.2698
Disgust	0.0071	0.0140	0.0093	0.2619
Happiness	0.0555	0.0891	0.0646	0.4365
Sadness	0.1982	0.2438	0.2354	0.1984
Fear	0.0378	0.0481	0.0431	0.1984
Surprise	0.0597	0.0801	0.0667	0.1984

The next consideration is the *spread of emotion values* for each frame of each emotion. One may notice in Table 3 that *neutral* again behaves as a benchmark, evenly distributing values with a relatively smooth and centred distribution. *Sadness*, the emotion closest to neutral, while somewhat skewed to the lower half of the distribution pentile shows an even distribution. All other emotions are far less evenly distributed, with most of the values compressed into the first and second pentile of values.

Table 3: Proportion of Values.

Emotions Analysed on 126 Frames 14 Clips with Edges B, G, H, I, K, M, N, P					
Emotion	Emotion Value Ranges				
	<0.2	<0.4	<0.6	<0.8	< 1.0
Neutral	0.054	0.302	0.339	0.229	0.076
Anger	0.959	0.036	0.004	0.002	0.0
Disgust	0.999	0.001	0.0	0.0	0.0
Happiness	0.977	0.010	0.004	0.004	0.005
Sadness	0.258	0.255	0.239	0.202	0.045
Fear	0.979	0.018	0.002	0.0	0.0
Surprise	0.954	0.033	0.007	0.006	0.0

Lastly, again consider the error data as seen in Table 2. Interpreting error requires the mean emotion values of any emotion to be high enough so that the region of the standard deviation is nearly all above zero. If the standard deviation region is clipped by a zero-value line, then the prediction values will likely rest above near-zero as well, providing no "bottom room" to dip below the standard deviation region. Unlike *sadness* as shown in Figure 6 and *neutral* seen in Figure 5, the predicted data for *scared*, *surprised*, *happiness* and *disgust* show unreliability for the NN and corpora for this research because their standard deviation regions drift over the zero-value line causing the PER and RMSE to appear to support accuracy, when in fact the NN is reacting with very little elicitation response for the given stimulus (flat

lining). Most interestingly, *sadness* (Figure 6) and to some degree, *anger* (Figure 7), show some promising responsiveness to the stimulus, reacting in similar ways as the mean of the frame of emotion values in the observed test data as indicated by the PER score for both in Table 2. With the RMSE score for *sadness* at 0.2438, its score is 0.0581 higher than neutral at .01857. *Anger* shows a lower RMSE, but anger values are still too low to be fully reliable with much of its standard deviation clipped by the zero line and the observed values of the test data also dropping to zero for nearly 20 frames.

The difference between the RMSE and the meanRMSE is that while the RMSE score looks only at the difference between the mean of the observed test data and the predicted data, the meanRMSE is the mean of all the plotted RMSE scores shown in red in each of Figures 4, 5 and 6. The plotted values show the RMSE for the chosen emotion at each of the 126 frames examined in relation to the same synchronized frame in the predicted value. The fact that the two RMSE scores are close in value, provides a check on the accuracy of the error assessment process.

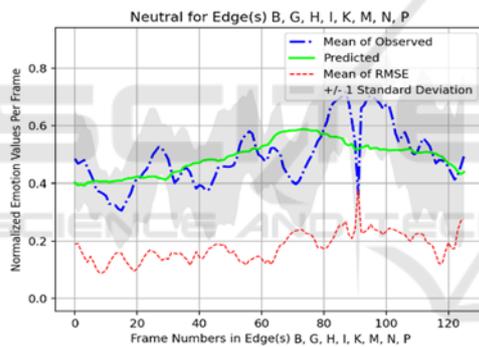


Figure 5: Neutral provides a benchmark for other emotions.

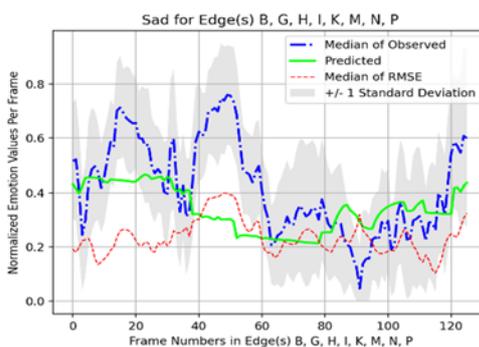


Figure 6: Mean of Observed values in midrange with standard deviation region unclipped.

One inexplicable anomaly is an apparent minimum value PER of 0.1984 in the test results shown in Table 2. Mathematically, it has been determined that for the

126 frame results for each emotion, 25 frames fell outside the standard deviation for the three emotions: *sadness*, *scared* and *surprise*. It remains unclear if this fact is a coincidence or caused by the test design.

6 CONCLUSION

Thus far, this paper has identified several statistical properties relative to *neutral*: RMSE, PAR, and Spread of Emotion Values. What has been demonstrated is that at least one emotion label, *sadness*, was successfully simulated. What might be useful is a classification system for each of these measurements that would provide discrete labels within a range. Such a classification system could indicate if the results will lead to a NN model that will output predicted emotion values that range from resemblant to not resemblant in relation to those values elicited by its human actor referent. The aim of this research is to expand the creative process of character design for video games beyond the modeller and animator and toward the skills of the actor. For the methods proposed to become useful, they must also produce salient results that confirm resemblance. Thus far, this research has demonstrated a statistical method to validate resemblance. Further investigation should confirm its viability as a method of production for game character production workflow.

REFERENCES

- Bates, J., Loyall, A. B., Reilly, W., (1994). An Architecture for Action, Emotion, and Social Behavior, In Artificial Social Systems: Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World, Springer-Verlag, Berlin.
- Barros, P., Churamani, N., Lakomkin, E., Siquiera, H., Sutherland, A., Wermter, S. (2018). The OMG-Emotion Behavior Dataset. In Proceedings of the International Joint Conference on Neural Networks.
- Bänziger, T., Mortillaro, M., Scherer, K.R., (2011). Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. In Emotion. vol. 12, no. 5. American Psychological Association, New York, NY, USA. 1161-1179.
- Busso, C., Burmanian, A., Sadoughi, N. (2017). MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. In Transactions on Affective Computing, vol.10, no. 10. 67-80. New York: IEEE.
- Cohn, J., Ambadar, Z., Ekman, P. (2007). Observer-Based Measurement of Facial Expression with the Facial Action Coding System, in Handbook of Emotion

- Elicitation and Assessment, eds. Coan, J. A., and Allen, J. B., Oxford University Press.
- Darwin, C., Prodger, P. (1872/1998). *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA.
- Ekman, P. (Ed.). (2006). *Darwin and Facial Expression: A Century of Research in Review*. Cambridge, MA: Malor Books, Institute for the Study of Human Knowledge.
- Khorrani, P., Le Paine, T., Brady, K., Dagli, C. and Huang, T.S., (2016). How Deep Neural Networks Can Improve Emotion Recognition on Video Data, in *IEEE International Conference on Image Processing 2016*, New York, NY, USA: IEEE, pp. 619-623.
- Kozasa, C, Fukutake, H., Notsu, H., Okada, Y., Nijima, K., (2006). Facial Animation Using Emotional Model, *International Conference on Computer Graphics, Imaging and Visualization*, pp. 428-433.
- Lewinski, P., Den Uyl, T. M., Butler, C. (2014). Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics* 7.4 (2014): 227.
- Loyall, A. B., (1997). *Believable Agents: Building Interactive Personalities* (No. CMU-CS-97-123), Carnegie-Mellon University, Department of Computer Science, accessed 12 October 2022 at: <https://www.cs.cmu.edu/afs/cs/project/oz/web/papers/CMU-CS-97-123.pdf>
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+). In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. pp. 94-101.
- Mascarenhas, S., Guimarães, M., Santos, P.A., Dias, J., Prada, R., Paiva, A., (2021). *FAtiMA Toolkit -Toward an Effective and Accessible Tool for the Development of Intelligent Virtual Agents and Social Robots.*, arXiv preprint arXiv:2103.03020.
- Metallinou, A., Lee, C., Busso, C., Carnicke, S., Narayanan, S. (2010). The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation. In *Proceedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*
- Moore, S. (1984). *The Stanislavski System: The Professional Training of an Actor*, Penguin Books, New York, NY, USA, pp. 41-46.
- Ortony, A., Clore, G. L., and Collins, A., (1990). *The Cognitive Structure of Emotions.* Cambridge, UK: Cambridge University Press, pp. 34-58.
- Paier, W., Hilsmann, A., and Eisert, P. (2021). Example-Based Facial Animation of Virtual Reality Avatars Using Auto-Regressive Neural Networks. *IEEE Computer Graphics and Applications*, 41(4), pp. 52-63.
- Paier, W., Hilsmann, A., and Eisert, P. (2020). Neural face models for example-based visual speech synthesis. In *European Conference on Visual Media Production*, pp. 1-10.
- Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3), 715-734.
- Schiffer, S., Zhang, S., Levine, M. (2022). Facial Emotion Expression Corpora for Training Game Character Neural Network Models. *VISIGRAPP*.
- Skiendziel, T., Rösch, A. G., Schultheiss, O.C. (2019). Assessing the Convergent Validity Between Noldus FaceReader 7 and Facial Action Coding System Scoring. In *PloS one* 14.10 (2019): e0223905.
- Soleymani, M., Larson, M., Pun, T., and Hanjalic, A. (2014). Corpus Development for Affective Video Indexing. In *IEEE Transactions on Multimedia*, 16(4), pp. 1075-1089.
- Suwajanakorn, S., Seitz, S. M., Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics*, 36(4), pp. 1-13.
- Vidal, A. Salman, A. Lin, W., Busso, C. (2020). MSP- Face Corpus: A Natural Audiovisual Emotional Database. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 397-405.
- Wingenbach, T., Ashwin, C., Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial expression set – Bath Intensity Variation (ADFES-BIV), In *PLoS ONE* 11(1): e0147112.