Measuring Emotion Velocity for Resemblance in Neural Network Facial Animation Controllers and Their Emotion Corpora

Sheldon Schiffer[a]
Department of Computer Science, Occidental College, 1600 Campus Road, Los Angeles, U.S.A.

Keywords: Facial Emotion Corpora, Emotion AI, Neural Networks, Non-Player Characters, Autonomous Agents.

Abstract: Single-actor facial emotion video corpora for training NN animation controllers allows for a workflow where game designers and actors can use their character performance training to significantly contribute to the authorial process of animated character behaviour. These efforts result in the creation of scripted and structured video samples for a corpus. But what are the measurable techniques to determine if a corpus sample collection or NN design adequately simulates an actor's character for creating autonomous emotion-derived animation? This study focuses on the expression velocity of the predictive data generated by a NN animation controller and compares it to the expression velocity recorded in the ground truth performance of the eliciting actor's test data. We analyse four targeted emotion labels to determine their statistical resemblance based on our proposed workflow and NN design. Our results show that statistical resemblance can be used to evaluate the accuracy of corpora and NN designs.

1 INTRODUCTION

As interactive entertainment continues to develop photo-realistic autonomous agents that behave with apparent emotional authenticity, a need has arisen to evaluate the accuracy of the methods used. Facial emotion corpora have been used to train NN controllers (Schiffer, 2022), but an evaluation method will help determine if a facial emotion video corpus produces sufficiently accurate data to train a NN. The design of the NN itself can also be a factor, though methods for evaluating NN efficacy is the topic for a separate investigation.

Two research communities are using facial emotion video corpora and are working toward similar ends. On the one hand, research in the Computer Graphics field have demonstrated many breakthroughs with the use of neural networks trained from live action video subjects to simulate facial elicitation. In most of these instances, the results have limitations of duration, emotional autonomy, and portability into multiple agents, but the demonstrated photorealistic visual qualities are highly accurate. On the other hand, researchers in the field of affective computing and computational psychology have shown great interest in simulating autonomous emotional behavior in virtual agents that elicit independently over lengthy durations. However, the resulting limitations of data transfer rates between NNs and the 3d meshes they control, demonstrate less precise graphic quality nor sufficient speed for commercial interactive animation, modeling, texture rendering and animation complexity.

Both research communities have deployed facial emotion corpora to train neural networks and used systematic workflow to produce a corpus of video samples. For both research communities, an evaluation method will assist in determining the quality of any individual video corpus and therefore the NN that used it for training. An evaluation method must determine if a video corpus can accurately train a NN to control the actuators of a Non-Player Character's (NPC's) facial elicitation system. The resulting animation of an avatar should be objectively similar to the actor's dynamic facial expressions while in-character and when provided identical stimulus within a virtual environment.

Velocity of emotion is one characteristic that describes the degree of intensity of the stimuli that triggered an elicitation, as well as the intensity of the emotional experience itself (Krumhuber et al., 2013). Therefore, quantifying and comparing velocity can

[a] https://orcid.org/0000-0001-5862-5239

Schiffer, S.
Measuring Emotion Velocity for Resemblance in Neural Network Facial Animation Controllers and Their Emotion Corpora.
DOI: 10.5220/0011676200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 1*, pages 240-248
ISBN: 978-989-758-623-1; ISSN: 2184-433X

demonstrate resemblance of reactive intensity in a NN-controlled avatar in relation to the human actor performing in a single-actor video corpus. If an NPCs elicitation dynamics express as the actor's character intended, narrative and rhetorical information in a video game or other interactive media can become clearer, enabling a story to unfold in a player's mind as intended.

Past research of video corpora has focused evaluation procedures on the classification of a corpus' video clips using surveys of human annotators. The primary intention of human-annotated corpus validation has been to warrantee that emotion label value assignments for static or dynamic images are valid for images of faces with random elicitations. These emotion labels and their recognition techniques evolved from Darwin's and Prodger's study of facial emotions in humans and animals (1872/1998) and seek to classify emotions through mostly static facial expressions as systematically practiced by psychologists Ekman and Friesen (1987) and Ekman (2006). Using these classification methods, facial emotion video corpora production and evaluation has contributed evidence that basic emotions are recognizable and classifiable. From this premise, NN design for facial emotion recognition has adopted a schema of emotion labels for emotion recognition widely known as the Facial Action Coding System (FACS), used to determine the degrees of elicitation of basic emotions observed through movements of muscle systems of the face (Cohn et al. 2007).

While the process of evaluating new facial emotion elicitation video corpora is important for research and the development of Facial Emotion Recognition systems (FERs), the interactive media industries are also burgeoning with similar needs for video corpora. The production of NPCs require an objective validation procedure integrated into a workflow to determine if the behaviors of an animated character are producing facial expressions as intended and as exuded by their human subject referent. Relying on human classifiers has value for creating FERs because they provide ground truth definitions of emotion elicitations for general-purpose emotion recognition systems. But for video corpora produced for training NNs to simulate an actor's performance through a photorealistic NPC, using human annotators is laborious and expensive.

The objective of this paper is to demonstrate a statistical method for researchers and developers to determine behavioral *resemblance* of a NN's behavior and the video corpora used to train it. Our method requires collaboration with actors and/or performance designers using familiar film and television performance preparation techniques to create a body of video recordings of facial emotion elicitations. This study uses a facial emotion elicitation corpus to train a NN facial animation controller. The NN is used to produce animation for a photorealistic avatar to compare the emotion *velocity* of an emotion label it produces with the emotion velocity elicited by a human actor in recorded video clips found in the corpora used to train the NN. Behavioral resemblance can be evaluated by isolating and measuring statistical characteristics of the emotion label values recognized by an FER of recorded facial expressions. We compare the velocity of data generated by an FER as it analyzes a live action stimulus source to that of predictive behavior generated by a NN animation controller as it reacts to the same stimulus source.

2 RELATED WORK

2.1 Example-Based Animation

During the last decade, the Computer Graphics community has developed new methods of simulating facial elicitation. This approach prioritizes graphical accuracy of modeling and animation, one expression at a time. Several studies by Paier et. al propose a "hybrid approach" that uses "example-based" video clips for frame-by-frame facial geometry modeling, texture capture and mapping, and motion capture (Paier et al., 2021). Their approach has shown remarkable graphic resemblance to a performing actor speaking a few lines or eliciting a series of pre-defined gestures. The approach of Paier et al. accomplishes graphical and performative resemblance by capturing face geometry for modeling, and dynamic facial textures for skinning a deformable mesh in real time (Paier et al., 2016). A NN using a variable auto-encoder (VAE) design is used to integrate motion for dynamic textures and mesh deformation, while another NN selects animation sequences from an annotated database to assemble movement sequences. Short single-word utterances or single-gesture elicitations are synthesized into sequences controlled by the developer. Database annotation of animation also demonstrates the efficacy of classifying movement data of a single actor that can be used later for semi-autonomous expressions. While the phrase "example-based" might be distinct from the word "corpus," the process of basing programmatic animation on data generated from a collection of videos depicting the same person is essentially a single-actor video corpus.

The emphasis on speech synchronization is found in the experiments of Paier et al. and a study by Suwajanakorn et al. that uses the vast collections of video samples of a U.S. president (2017). From a 17-hour corpus of President Obama, the investigators mapped speech from persons who were not Obama, onto a moving and speaking face of the presidential subject. Their method discovered that training their NN benefited from considering 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 utterance. 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 our 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 provide a sufficient model of fully autonomous elicitation in response to measurable aleatory stimuli (e.g., the interactive face of a human user, player, or a non-interactive face from an NPC). Autonomous emotional agent design has relied on models developed in Computational Psychology over decades of research. These most current approaches of the graphics community do not classify facial expressions using FACS or other widely acknowledged emotion interpretation system. No classification system of emotions for their video corpora are disclosed, and therefore the meaningfulness of synthesized expression relies on arbitrarily selected spoken semantics and correlating visemes. While the workflow design of both Paier et. al and Suwajanakorn et al. are impressive for their mimetic capacity, their fundamental experimental design provides little computational means to control emotion as a quantity itself, but only facial expression as an elicited instance detached from an internal psychological cause.

2.2 Facial Emotion Velocity

Previous corpora production deployed for the development of Facial Emotion Recognition (FER) systems has validated its research using annotators' judgement of mostly *static* video frames. Relatively little attention has been given to the perception of velocity of facial emotion expression. Research has shown that facial expression in motion provides more recognition efficacy than static expression recognition (Krumhuber et al., 2013). Further research suggests that perceptions of "naturalness" were greatly affected by changes in expression velocity (Sato and Yoshikawa, 2004). If we also examine the same behavior across disciplines, we find that emotion expression velocity, often called tempo or rhythm in the performing arts, is a physical manifestation of a character's inner state in relation to the outer circumstances (Morris, 2014). Some emotion expressions change in tempo as determined by the intensity of the emotion and affected by the "inner needs" of a character, and the physical conditions of the character's circumstance. While the performing arts community has for many decades acknowledged the importance of elicitation velocity, measuring changes in emotion labels over time has not been prioritized as a measurable emotional property in facial elicitation corpora annotation. To measure modulations in emotional velocity in corpora, the video sample production method must consider dynamic stimuli. For our proposed method, we adapted a technique of recording multiple versions of the same action with different emotional intensities to trigger varying emotional velocities.

2.3 Corpora Production

Emotion recognition video corpora provide a ground truth baseline for general emotion recognition. 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) and Benda et al., (2020) versus non-actors (Vidal et al., 2020). Similarly opposed approaches is the choice to use tightly scripted scenarios as did Busso et al. (2017), versus more improvisational techniques (Metallinou et al., 2010). Lucey et al. proposed to record video in a lab-controlled environment (2010), while others collected samples from major media production industries such as news and entertainment (Barros et al., 2018). Our approach was to use actors with a scripted scenario written to generate specific emotions for a scripted video game scene with varying paths.

The evaluation methods we developed required fewer elicitation variations than a corpus and NN designed for generic emotion recognition. Nonetheless, the validation methods used for FERs provide insight for our approach. Most involved either a process of soliciting human observers to annotate video clips, or they require participating subjects to self-annotate classifications of their own elicitations (Soleymani et al., 2014). Nearly all corpora reference the Facial Action Coding System (FACS), as does our research. FACS correlates groups of muscles, called action units (AUs), that manipulate the face to form expressions of at least six basic emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* (Cohn et al., 2007). Classification methods solicit an annotator to identify an emotion label in a still image of a video clip. They estimate its intensity or provide perceived levels of arousal and valence (Soleymani et al., 2014). This approach provides a ground truth reference to be validated with statistical evidence for FER performance evaluation. To implement human annotation for NN-generated animation in an NPC would duplicate the work of the FER system used to analyze the original corpora clips. Furthermore, the process of collecting human annotation would defeat the interest of saving labor costs and production time. Therefore, we elected to use FER systems for annotation of all clips without testing for human-machine inner-annotation agreement. Automated classification has been shown to correlate accuracy with human classification for dynamic expressions (Calvo et al., 2018).

Most corpora segregate statistical tallies by emotion label, valence, or arousal. These classifications can then be further discerned by observations of intensity and elicitation duration. 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 mesh modeled after the performing actor of the corpus. This approach allows for classifying resemblance by evaluating the difference between the emotion label values of the performing actor and their avatar, and the difference in change of value over time, which is the emotion label value velocity.

2.4 Neutral Network Design for NPCs

In addition to corpus development, NN development also has a history of related work. 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 for actors in-training to implement on virtual agents in digital and interactive media (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 and Collins, also known as Appraisal Theory (Ortony et al., 1990). Appraisal Theory provides a structured model of emotion generation that resembles many of the features of Stanislavski's Method approach, where emotion is the result of events whose outcomes may help or hinder one's psychological and physical needs. Appraisal Theory was intentonally organized as an architecture for computational simulation of the human emotion system. These approaches from psychology and the performing arts were implemented into code in the virtual agents of the Oz Project experiments (Bates et al., 1994). Loyall et. al produced several interactive media experiments with autonomously animated characters using emotion AI systems that deployed digital models of emotion.

Implentation architectures have continued to progress for autonmous emotion elicitation systems. Kozasa et. al. showed an early use of an affective model for an emotive facial system in an NPC based on a dataset of expressions (2006). Theirs used a 3-layer feed-forward artificial neural network to train an NPC from "invented" data for parameters fed to a NN model. They claimed no databases at the time existed to train their model. Later, using appraisal theory-based design from virtual agents, the FAtiMA architecture was integrated by Mascarenhas et al. with a NN model in educational games (2021). In its earlier versions in social and educational games, the FAtiMA architecture had proven to be effective for learning and engagement (Lim, et al., 2012). Khorrami et. al showed that the use of LSTM cells for emotion recognition of facial video was shown to improve previous NN performance for emotion recognition (2016). Unlike the method that this research proposes, these previous related works did not use single-actor video corpora.

3 METHODS

3.1 System Overview

The proposed method combines a series of steps that require skilled participants with expertise in distinct disciplines. These include video game scenario writing, performance rehearsal, video production, facial emotion analysis, NN design, 3d facial modelling, 3d animation and data analysis. Since game design is not a linear narrative form, we elected to create a scripted dialog-behaviour tree authored in the form of a directional acyclic graph as seen in Figure 1. Any path through the tree can be performed as a script. The tree allows for a calculable number of distinct paths that we rehearsed and recorded to video clips to create facial emotion corpus.

The video clips of a respondent human Emotion Model and those of their human Stimulus Source were analyzed using the *Noldus FaceReader 8* FER system. The FER-generated data from resulting video corpus were the basis for producing a corpus of clips from which to train a NN.

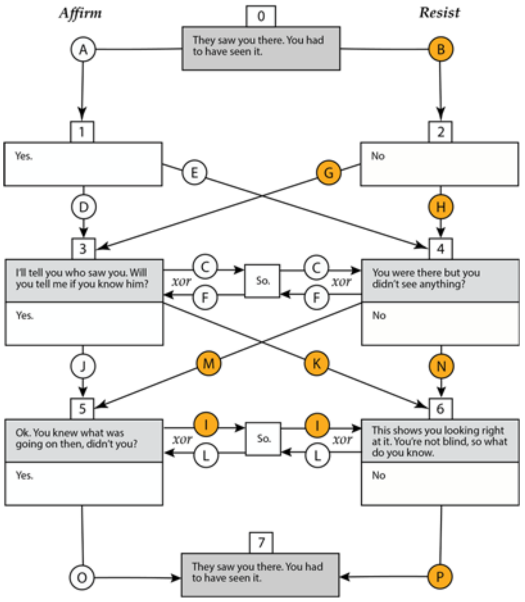

Affirm
Resist
0
They saw you there. You had to have seen it.
A
B
1
2
Yes.
No
E
G
D
H
3
4
I'll tell you who saw you. Will you tell me if you know him?
Yes.
C
F
xor
So.
C
F
xor
You were there but you didn't see anything?
No
J
M
K
N
5
6
Ok. You knew what was going on then, didn't you?
Yes.
I
L
xor
So.
I
L
xor
This shows you looking right at it. You're not blind, so what do you know.
No
7
O
P
They saw you there. You had to have seen it.

Figure 1: Circled labels are edge segments. Orange-colored were used in the experiment. As an acyclic graph, a path can use edge segment C *xor* F and I *xor* L.

3.2 Producing the Corpus as a Tree

In our experiment, the scenario was rehearsed in advance of production with interaction between two characters, one an Emotion Model and the other, an emotion Stimulus Source. Actors were rehearsed based on a backstory where the Emotion Model is being interrogated by an investigator about a crime. Actors were asked to focus on mental actions, such as *affirm* or *resist*, that could be performed with few words and facial expressions. The scenario as depicted in Figure 1, has 32 possible paths to resolution. The Stimulus Source performs the role of an interrogator investigating a crime asking the Emotion Model accusatory questions. The Stimulus Source sat in front of a camera in a close conversational position, and was pre-recorded and edited to consistent lengths such that all possible paths of the dialog tree could be presented with edge segments cut to consistent lengths and the human Emotion Model could respond synchronously with the video recording of the Stimulus Source as illustrated in Figure 2. The Emotion Model performed the role of suspect. They watched and reacted to each of the 32 possible paths of video interrogation as if the Stimulus source was talking and eliciting in person. The dialog was written so that the Emotion Model could only respond with the words "Yes," "No" or "Maybe." The resulting facial emotion corpus consisted of recorded clips of the Emotion Model's performance of each of the 32 paths. Each path was recorded 9 times in triplets. For each of the triplets of clips, the actor was given cause and direction to express three distinct degrees of intensity so that each triplet would be either low, medium, or high in intensity, thus allowing for distinct intensity and velocity modulations between each of the triplets. Thus, 9 clips for each of 32 paths yielded 288 clips of lengths ranging from 40 to 50 seconds.

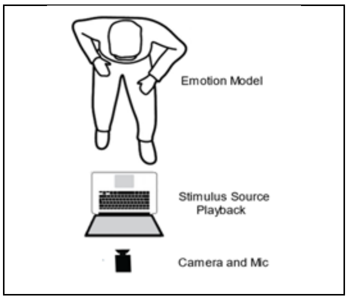
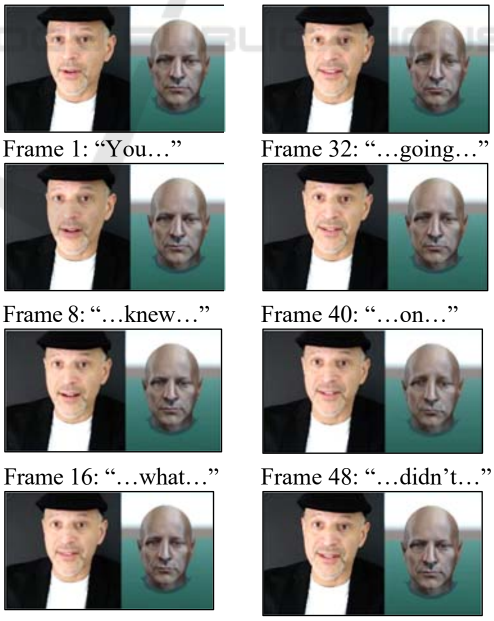


Figure 2: Setup for sample recording.

3.3 Modelling the Avatar

To create an avatar of the Emotion Model, *FaceBuilder* was used for 3d head modeling and animation. It automatically creates 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



Frame 1: "You…"

Frame 32: "…going…"

Frame 8: "…knew…"

Frame 40: "…on…"

Frame 16: "…what…"

Frame 48: "…didn't…"

Frame 24: "…was…"

Frame 56: "…you?"

Figure 3: Eight synchronized frames over 2.3 seconds show Stimulus Source (left) providing action: *accuse*. NN Avatar (right) reacts with sadness.

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. Blend shapes are game engine acturator that can receive streamed emotion data from the NN animation controller responding to a human Stimulus Source's face that performs as the game player. An embedded NN receives FER data and controls the avatar of the human Emotion Model to "react" in a way that intends to statistically resemble the character behavior the actor created in the video clips of the single-actor corpus (See Figure 3).

The NN-generates its predictive data as the reaction in the form of normalized emotion label values as a means of autonmously controlling the FaceBuilder head mesh to animate facial expressions. These values of the NN-controlled Avatar were compared for their changes over time during synchronized instances of stimuli from the human Stimulus Source.

3.4 Modelling the NN

The NN was trained from the clips produced for the single actor corpus. 68.8% (198 clips) were used only for training the NN. 20.1% (58 clips) were used only for validation of the NN. The remaining 11.1% (32 clips) was used to test the NN's behavioral resemblance to the actor's character. Test clips traversed 4 paths with 14 edge segments in common that consisted of 32 individual clips of between 40-60 seconds length. These clips yielded at least 3840 frame instances of emotion label value data to compute emotion velocity and test behavioral resemblance.

The principal components of the NN follow a Recurrent Neural Network (RNN) design. Predicting the facial elicitation of game characters based on training data from an actor's performance requires both spatial and temporal data representation. 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. Facial feature positions were probabilistically estimated from their spatial contexts using Time-Distributed Dense architectures. The experiment of this research used a Dense layer of perceptrons that are were connected to two layers of 100 bi-directional LSTM cells. The LSTM layers auto-regressively consider data from 10 seconds in the past. Since the data for this experiment was fed preprocessed emotion data tables (as opposed to a live video stream), the NN also analyzed 10 seconds into the future.

Each emotion label was assigned its own NN, so the recurrent NN was cloned into an entourage of 7 different NNs. For the experiment, the developed game scenario anticipated and targeted four probable resemblant emotions: *anger*, *fear*, *sadness* and *surprise* each with their own NN. By separating the emotion labels into distinct NNs, we were able to observe which NN was most resemblant for velocity.

3.5 Post-Processing Emotion Analysis

To create the emotion dataset of the corpus used to train the NN, dynamic facial expressions of each clip were post-processed by a Facial Emotion Recognition (FER) software. *Noldus FaceReader 8* generated normalized emotion label values (0.0 to 1.0) of four targeted basic emotion labels: *anger*, *fear*, *sadness*, *surprise*. Noldus FaceReader 8 is a tested and ranked FER system that has produced emotion recognition validation results that match the accuracy of human annotators (Lewinski et al. 2014). Furthermore, the recognition accuracy rate of FaceReader has been documented as high at 94% (Skiendziel et al. 2019).

FaceReader 8 continuously recognized the four targeted emotion label values for 3 frame instances per second of video. The corpus consisted of 288 video clips where each of the 32 paths of the dialog behavior tree were perfomed 9 times creating three triplets, each with low, medium and high intensity emotion responses motivated by changes in the backstory of the video game scenario written for the dialog behavior tree.

3.6 Evaluation Method

The video clips were subdivided by synchronized frames that share timecode with clips of the Emotion Model and the Stimulus Source. Each frame correlates to the individual edge segments of the 32 possible dialog behavior tree paths. For the experiment and results presented in this paper, we examined a set of 9 edge segments in common among 3 paths: nodes 0-2-4-5-6-7, 0-2-4-6-7 and 0-2-3-6-7, amounting to 27 clips for the experiment. Each of these edge segments represent a total of 1008 frames or 42 seconds of video. We used only the orange colored segments from Figure 1 (B, G, H, I, K, M, N, P). For each of the three paths and their 9 video clips, the emotion data of four targeted emotions: surprise, anger, fear and sadness was taken from the first and last frame of the first third, second third, and last third (14 seconds) of each their entire 42 second paths. The sub-division of the clips allowed for shorter duration of equal lengths to give more accurate velocity means for each analyzed sub-clip. Each of the sub-clips consist of 336 frames or 14 seconds. FaceReader 8 was set to analyze emotions for one frame of video for every 8 frames, which is the equivalent of 3 fps out of 24 fps. Therefore each segment is analyzed 42 times (336 live action frames ÷ 8 FER analyzed frames per second).

To calculate mean velocity for each emotion across each 336-frame segment, we first found the mean emotion e_μ at each frame for each path from all clips that traverse the same edge segment using the following equation.

$$e_\mu = \frac{\sum_{S=1}^{S} e_k}{S} \qquad (1)$$

The quotient e_μ is found by dividing the sum of values e at frame k for each emotion. The set S consists of all emotions e_k that occur at the same analyzed frame. There is an e_μ for every analyzed frame in every path shown in

Figure 1. With e_{μ} found for all analyzed frames in the experiment, we compute the velocities of each sub-segment of 336 frames containing 42 analyzed frames as follows for v_{μ}.

$$v_{\mu} = \frac{e_{\mu t41} - e_{\mu t0}}{t_{41} - t_{0}} \tag{2}$$

Each of the velocity values were computed as the difference of emotion values at time 0 and time 41 for each of the 81 sub-clips divided by the difference in time from the first analyzed frame of the 14 second sub-clip to its last analyzed frame.

For both the NN-generated predicted data that controled the Avatar, and for the FER generated data from the Emotion Model, we aggregated all mean velocity values of each emotion to find a standard deviation, mean and variance. Since all prediction data that animated the NN-controlled Avatar synchronously aligned frame-for-frame with the Stimulus Source video clips as positioned by their dialog behavior tree segment, emotional label values and their velocities corresponded to the emotion values at frames triggered by the behaviors of the face of the Stimulus Source. The statistical comparison of the predicted data and test data is revealed in Results.

4 RESULTS

Figures 4, 5, 6 and 7 show 81 velocity means of sub-clips distributed in histograms for four targeted emotions. The four NNs performed similarly. Most notably is the narrow variance in the histograms. The widest variance occurs with *sadness* at 3.47e-5, which is a magnitude of 2 less than its mean velocity. (See Figure 4.) These narrow variance results indicate that the Avatar emotion velocity is behaving similarly to the Emotion Model. We note in all cases, the mean velocities of the four targeted emotions frequently show that the Avatar responded with less velocity than the Emotion model. The concentration of low velocity reactions to near zero is greater for the Avatar than the Emotion Model. We notice that as the distribution of velocity disperses away from zero and toward the extreme limits of the range, we find the number of edge segments with accelerating or decelerating velocity to be similar. Outside of the main distribution curve that surrounds either side of zero, the outlying velocity means in the range fall on different values. But their count of edge segments differs between the data sets by less than 5. The outliers outside the normal curve around zero suggest the Avatar chooses a different intensity of response to the Stimulus Source than does the Emotion Model.

Since the objective of this research is to determine if the NN design and the video corpus show sufficient resemblance in emotion velocity behavior, we chose to see if the mean of the Avatar emotion velocities falls within ± 1 Standard Deviation of the Emotion Model velocities. Table 1 shows that for all four emotion labels, this requirement is met. We notice that the Avatar mean always falls on the negative side of the Emotion Model's velocity distribution. This suggests that the Avatar starts each segment with a higher value than when it ends, and that there is decline in intensity over time that the Emotion Model does not exhibit.

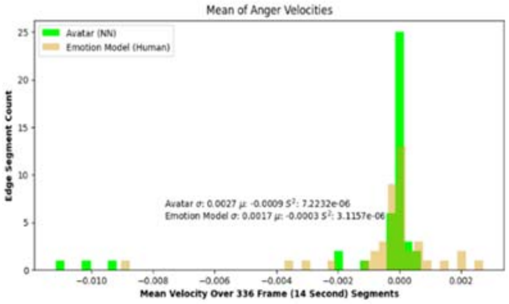


Figure 4: Distribution of Anger Velocities.

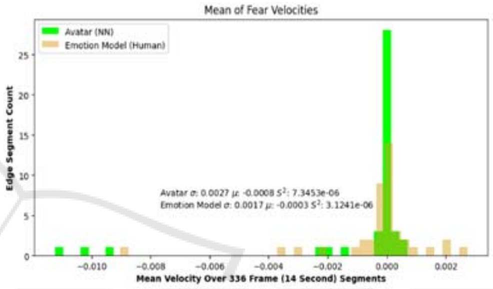


Figure 5: Distribution of Fear Velocities.

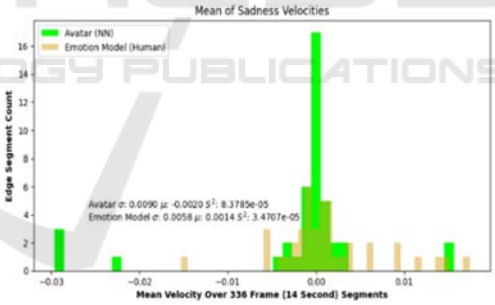


Figure 6: Distribution of Sadness Velocities.

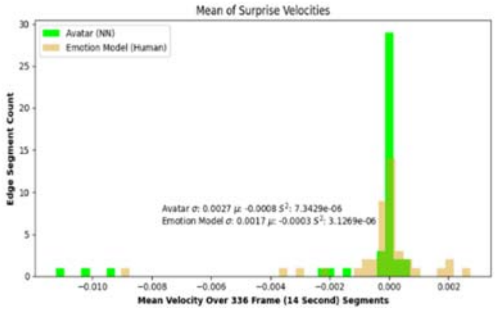


Figure 7: Distribution of Surprise Velocities.

Table 1: Mean Velocities Over 81 Edge Segments.

Emotion	Mean Avatar	Mean Em.M.	Abs. Diff.	S.D. Avatar	SD Em.M.
Anger	-.0003	-.0009	.0006	.0027	.0017
Fear	-.0003	-.0008	.0005	.0027	.0017
Sadness	.0014	-.0002	.0016	.0090	.0058
Surprise	-.0003	-.0008	.0005	.0027	.0017

5 CONCLUSION

Locating the mean of the Avatar velocities for four emotion labels within ± 1 Standard Deviation suggests that the Avatar's velocity behavior is within at least 68% of the behavior of the Emotion Model's behavior in response to the same stimuli. This fact alone must be bolstered by noting that the Variance for each of the histograms is 1 to 2 magnitudes less than the Standard Deviation. With such a narrow variance, we conclude that the emotion velocities of all four emotion labels are *substantially resemblant*.

The contribution or our proposed method is to create a ground truth reference for one single subject. The experiments of this research accept the assumption that the corpora used to validate the NNs embedded in the FER software are sufficient to build a secondary corpus like the one we propose, designed to simulate one actor's character rather than to recognize the facial emotion of a wide set of generic human faces. The approach of this research intends to streamline character animation in video game production by leveraging the work of human annotators used in the development of FERs. By using programable statistical techniques as applied in this research, a more automatic process of evaluating facial emotion corpora could accelerate the use of emotion AI in NPCs for future game production.

REFERENCES

Bänziger, T.; Mortillaro, M.; and Scherer, K.R., (2011). Introducing the Geneva Multimodal Expression Corpus for Experimental Re-search on Emotion Perception. In Emotion. vol. 12, no. 5. American Psychological Association, New York, NY, USA.

Bates, J., Loyall, A. B., and 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.; and Wermter. S. (2018). The OMG-Emotion Behavior Dataset. In Proceedings of the International Joint Conference on Neural Networks (IJCNN).

Benda, M. S.; Scherf, K. S. (2020). The Complex Emotion Expression Database: A validated stimulus set of trained actors. In PloS One, 15(2), e0228248

Busso, C.; Burmania, 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. New York: IEEE

Calvo, M. G., Fernández-Martín, A., Recio, G. and Lundqvist, D. (2018). Human observers and automated assessment of dynamic emotional facial expressions: KDEF-dyn database validation. In Frontiers in Psychology.

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., and 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.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., and Tzavaras, A. (1987) Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology, 53(4), 712–717

Khorrami, 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 (ICIP), New York, NY, USA: IEEE.

Kozasa, C, Fukutake, H., Notsu, H., Okada, Y., and Niijima, K., (2006) Facial animation using emotional model, International Conference on Computer Graphics, Imaging and Visualization (CGIV'06).

Krumhuber, E. G., Kappas, A., and Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. Emotion Review, 5(1).

Krumhuber, E. G.; Skora, L.; Küster, D.; Fou, L. (2017). A review of dynamic datasets for facial expression research. In Emotion Re-view, 9(3).

Lewinski, P., Den Uyl, T. M., and Butler, C. (2014). Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader. Journal of Neuroscience, Psychology, and Economics 7.4.

Lim, M.Y., Dias, J., Aylett, R., Paiva, A., "Creating adaptive affective autonomous NPCs," Autonomous Agents and Multi-Agent Systems, 2012, New York, NY, USA: Springer.

Loyall, A. B., "Believable agents: Building interactive personalities" Carnegie-Mellon University, Pittsburgh, PA Department of Computer Science, (1997) accessed 30 June 2021 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.: and Matthews, I. (2010) The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE.

Mascarenhas, S., Guimarães, M., Santos; P.A., Dias, J., Prada, R. and 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.; and Narayanan, S. (2010) The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation. In Proceedings of Multimodal Corpo-ra (MMC 2010): Advances in Capturing, Coding and Analyzing Multimodality

Moore, S. (1984) The Stanislavski System: The Professional Training of an Actor, Digested from the Teachings of Konstantin S. Stanislavsky, Penguin Books, New York, NY, USA.

Morris, E. (2014). The ins and outs of tempo-rhythm. Stanislavski Studies, 2(2).

Ortony, A., Clore, G. L., and Collins, A., “The Cognitive Structure of Emotions,” Cambridge, UK: Cambridge University Press, 1990, 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).

Paier, W., Kettern, M., Hilsmann, A., and Eisert, P. (2016). A hybrid approach for facial performance analysis and editing. IEEE Transactions on Circuits and Systems for Video Technology, 27(4).

Paier, W., Hilsmann, A., and Eisert, P. (2020, December). Neural face models for example-based visual speech synthesis. In European Conference on Visual Media Production.

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).

Sato, W., and Yoshikawa, S. (2004). Brief Report: The Dynamic Aspects of Emotional Facial Expressions. Cognition and Emotion, 18(5).

Schiffer, S.; Zhang, S. and Levine, M. (2022). Facial Emotion Expression Corpora for Training Game Character Neural Network Models. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – HUCAPP.

Skiendziel, T., Rösch, A. G., and Schultheiss, O.C. (2019) Assessing the Convergent Validity Between Noldus FaceReader 7 and Facial Action Coding System Scoring. 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).

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4).

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 (ICMI ‘20). October 2020. 397-405.