Facial Emotion Expression Corpora for Training Game Character Neural Network Models

Sheldon Schiffer[1], Samantha Zhang[2] and Max Levine[3]
[1]Department of Computer Science, Occidental College, Los Angeles, California, U.S.A.
[2]Department of Computer Science Cornell, University Ithaca, NY, U.S.A.
[3]Department of Computer Science, University of North Carolina Asheville, Asheville, NC, U.S.A.

Keywords: Facial Emotion Corpora, Video Game Workflow, Non-player Characters, Video Games, Affective Computing, Emotion AI, NPCs, Neural Networks.

Abstract: The emergence of photorealistic and cinematic non-player character (NPC) animation presents new challenges for video game developers. Game player expectations of cinematic acting styles bring a more sophisticated aesthetic in the representation of social interaction. New methods can streamline workflow by integrating actor-driven character design into the development of game character AI and animation. A workflow that tracks actor performance to final neural network (NN) design depends on a rigorous method of producing single-actor video corpora from which to train emotion AI NN models. While numerous video corpora have been developed to study emotion elicitation of the face from which to test theoretical models and train neural networks to recognize emotion, developing single-actor corpora to train NNs of NPCs in video games is uncommon. A class of facial emotion recognition (FER) products have enabled production of single-actor video corpora that use emotion analysis data. This paper introduces a single-actor game character corpora workflow for game character developers. The proposed method uses a single actor video corpus and dataset with the intent to train and implement a NN in an off-the-shelf video game engine for facial animation of an NPC. The efficacy of using a NN-driven animation controller has already been demonstrated (Schiffer, 2021, Kozasa et. al 2006). This paper focuses on using a single-actor video corpus for the purpose of training a NN-driven animation controller.

1 INTRODUCTION

The stated goals of various academic emotion corpora have largely focused on creating multi-modal recordings for data acquisition to test theories of human emotion generation and inter-agent trans-cultural emotion literacy. Also, many corpora datasets are designed to train neural networks (NNs) for emotion recognition. The fundamental premise is that with enough diversely designed corpora and datasets generated from video samples, a generalizable framework of understanding about emotion would emerge from rigorous experimentation and data analyses. Underlying this assumption is that the behaviour of the subjects in the video samples can be validated for sufficient degrees of "authenticity" and "naturalness". The resulting research contributes to NN systems for facial emotion recognition (FER) and to facial emotion simulation in virtual agents, including video game characters. The entertainment industries (primarily video game, film, and television) are positioned to utilize this research in ways that other industries have thus far mostly ignored. While agencies in government, military, security, and product marketing have a great interest in predicting the thoughts, emotions, and behaviours of large sets of randomly selected "real" people, the entertainment industries make their business by creating and predicting the thoughts and behaviours of synthetic persons – fictional lead characters in what has become a variety of transmedia metaverses and their characters. This research proposes developing video corpora of one actor to train NNs that animate the facial expressions of photorealistic NPCs that closely resemble that specific actor.

Capturing volumetric and motion data of individual actors in-character is already a common task of computer-generated asset creation. A larger challenge is animating computer-generated characters such that movement is autonomous and

Schiffer, S., Zhang, S. and Levine, M.
Facial Emotion Expression Corpora for Training Game Character Neural Network Models.
DOI: 10.5220/0010874700003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 2: HUCAPP*, pages 197-208
ISBN: 978-989-758-555-5; ISSN: 2184-4321

algorithmically controlled by an AI system and not by linearly determined animation paths. Just as important, resulting movement should retain the signature gestural features of the individual actor performing as the character, and not prototype movement implemented from a motion capture corpus of linearly recorded animations.

Conventionally, game engines provide predetermined conditions that, when triggered, the movement fragments execute systematically through an automated animation controller, often a stack-based finite state machine (SBFSM). But to create natural movement with a SBFSM is to allow for many more stored sub-fragments of movement data activated within nested and parallel SBFSMs chained in complex sequences. Theoretically, this would require a logarithmic expansion of the possible combinations of these movements. But even with more short-term memory made available and skilful use of SBFSMs, where each state would remove itself from short term memory when obsolete, managing the memory stacks of multiple simultaneously engaged nested and chained SBFSMs creates excessively complex memory management routines such that game system processors become overwhelmed with processing too many states of character animation. Photorealistic cinematic character performance aesthetics requires more efficient solutions for game developers.

This research proposes two novel steps toward advancing autonomous facial expression animation. First, the implementation of an actor-centric performance-theory-informed method to create a corpus of video samples that captures actor-specific expression. Second, a novel way to validate the corpus using off-the-shelf FER system software and statistical analysis of emotion data generated by the FER system. The remainder of the paper is divided as follows: Section 2 examines the production and validation methods of a variety of recently produced facial emotion corpora. Section 3 describes the production methods used to create the facial emotion corpus for this research. Section 4 describes the techniques used to validate the corpus and the statistical results they produced. Section 5 provides a discussion of open questions about corpora production and what future directions this research may take. It also offers general suggestions for other investigators to consider if similar techniques were to be deployed. Section 6 concludes by evaluating what this research offers for implementation in the broader field of NPC animation in video games.

2 RELATED WORK

The experimental design of this paper was drawn from two principal areas. First, this investigation considers the production process of previous corpora that used actors as subjects for elicitation recording. Second, this investigation examines and adapts character development and rehearsal techniques drawn from acting and performance theory. This paper will summarize findings related to acted corpora and some "natural" corpora where "real" people are used as behavioural subjects. Additionally, deployable acting theories for corpus production will be discussed.

Multimodal emotion corpora have been categorized by a variety of properties that describe their production and therefore their utility. A first consideration is if a corpus was acted and produced by the investigators in a controlled environment or were its video samples gathered from the billions of clips available on the many user-contributed video-streaming websites. For the case of using actors, this research considers if the video samples portrayed actors as characters eliciting an emotion, or instead showed actors performing as themselves but induced into an emotional state, using what is widely known in acting theory practiced in 20th century United States academies as inducing an affective memory (Moore, 1984). Additionally, in the case that actors performed as characters, were the elicitations induced using another technique known as substitution (Strasberg, 2010), or was the emotion produced from a dyadic conversation using improvisation principles of active analysis technique (Thomas, 2016). As this paper will reveal, the expository rationales published with the release of the corpora reviewed here, hint at the acting methodology used, and describe variations of acting methods that give credit to the performance theorist whose ideas the affective computing community owes significant reference. Before expanding further on the methods that this paper and its experiment advances, it is important to address the types of corpora production which the experiment of this paper considered.

A prevailing trend for the development of emotion corpora is the use of "in-the-wild" or "natural" footage, most often found in the databases of user-provided video streaming websites. These corpora depict "real" people exuding a variety of emotional states in a wide range of contexts. This paper and its experiment considered if the on-screen subjects were aware of the camera's presence or were they caught in an emotive state unaware of the camera. Another consideration was if the corpora sample production or

sample collection demonstrated consideration of the context of a recorded elicitation, or was the sample collected as a decontextualized fragment. Included in the contextual concern was the presence of other persons, their relationship to the elicitor, if the person behind the camera was acquainted with the subject, and if the stimuli that triggered the emotion was present in the frame, off-screen, or out of the scene altogether. Unfortunately, most of these concerns were not consistently answered in the published rationales for the creation of corpora with "natural" or "lay" elicitors. Yet neuroscientific research confirms the hypothesis that social context can inhibit emotional elicitation (Clark, 1996) largely through what has been called a neurologically "constructed emotion" (Feldman Barrett, 2017). Furthermore, the affective computing community has remarked on the ethical concerns (Scherer and Bänziger, 2018) and the accuracy of using such footage for corpora development to study or model emotion due to inhibition of emotion display (Scherer, 2013). Some critics have also questioned the accuracy of "in-the-wild" corpora that rely on instantaneous emotion evaluations and ignore context and self-regulating temporal elicitation (Barros et. al 2018). This paper concludes that the corpus sample set to use for modelling a single character should consist of elicitations whose context (imagined by the actor) approximately matches those of a future context (perceived by the video game player) from which the data intends to inform (for classification) or predict (for NN modelling). The input emotion should categorically resemble that of the predicted output.

Research for this paper sought to find corpora created for the same purpose as its hypothesis. No similar multimodal or unimodal corpora was found with the intent to train a NN for a single-NPC. But what follows is a summary of rationale that highlights the suitability of using actors to create multimodal corpora for the general study of emotion and to validate some specific concepts about emotion. These rationales also remark on the pitfalls of using lay elicitors. The principal corpora that advocated the use of actors is the Geneva Multimodal Expression Corpus and its emotion portrayals Core Set (GEMEP and GEMEP Core Set). Bänziger et. al note the advantage of "experimental control" and "validity of the stimulus" that reveal the importance of understanding the context and the system of cues that create, encode, and decode an emotion elicitation (Bänziger, Mortillaro, Scherer, 2011). Recording samples can be acquired in a "holistic fashion" without actors being aware of the recording process. Rather than ask for disconnected elicitations, facial movement and speech can come because of a "specific episode often characterized by a scenario" (Busso et. al 2008). Corpora production can deploy what Bänziger et. al call "felt experience enacting," where an induction method based on remembered events and imagery cause emotions to be recalled sufficiently to elicit facial expression. The results reported for the techniques applied showed "significant expressive variations rather than a common prototypical pattern" (Busso et. al 2008).

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP, 2007) likewise highlights similar methods including the use of "experienced actors" speaking "natural dialogues in which the emotions are suitable and naturally elicited." The recommendation from Busso et. al is to record the samples emphasizing control over emotional and linguistic content, suggesting that emotions should be "targeted", and dialog should be performed from memory, with less improvisation (Busso, et. al, 2017).

The MSP-IMPROV database also involved producing its corpus with actors. While utilizing many of the same principles established in the two previous corpora mentioned, MSP-IMPROV developed a unique emphasis on fixing the lexical content while recording variations on how it can be elicited (Busso et. al 2008). Lexically identical sentences were recorded with the actor creating distinct scenarios for each recorded version, causing different emotions to surface for the same set of words. The intention was to discover how and if emotions surfaced in voice but not through the face, and vice-versa, when such variations of expression are designed to occur. While not mentioned in the published rationale for the corpus, this technique was earlier used in training American actors by Sanford Meisner. The technique is widely known as the Repetition Exercise in acting communities and the exercise has many variations still practiced by Meisner technique instructors (Meisner, 1987).

Complimentary in its purpose and approach, the CreativeIT corpus also used improvisational technique that yielded statistically significant consistency among its annotators. Most useful to the experiments deployed for this paper is the deployment of Stanislavsky's active analysis technique, where actors in dyadic conversational mode focus on a verb-action word to accomplish a goal that the other actor controls. The CreativeIT shared one goal of this paper and its experiments. It was the only corpora that explicitly set out to create a corpus that would assist in analysing "theatrical performance" for entertainment in addition to the

broader application of "human communication" (Metallinou et. al, 2010).

The One-Minute Gradual-Emotion Recognition (OMG-Emotion) dataset is among the corpora that used lay elicitors from videos harvested on a streaming website. The published exposition of the corpus critiqued previous corpora that used acted emotion elicitation (Barros, 2018). Barros et. al found that too much focus on controlling lexical content may have provided consistent and discrete evaluations of recorded utterances for targeted emotions, but that too often acted corpora lacked long enough or diverse enough expression to evaluate transitions from one emotional state to the next. OMG-Emotion focused its harvesting of multimodal samples on providing more complete context for shifts that showed how a subject changed over time. The investigators' emphasis on analysing context can be applied to acted corpus sample production provided the sample and the rehearsal have sufficient duration.

This paper found useful the MSP-Face Corpus (Vidal, Salman, Lin and Busso, 2020), another of the corpora consisting of video-sharing content. Its production methods were completely different from the experiment of this paper. It does not consist of acted elicitations but instead is a collection of recordings of "real" people. Its samples are cut into small segments that can be described with one global emotion descriptor. The collection was curated to focus on subjects in a front-facing position. The result of the curatorial focus on consistent temporal and positional requirements of each subject and sample appears to have resulted in a relatively even distribution of annotation agreement among samples where the dominant emotion label out of eight is not more than 23% of the collection. Such an even distribution is a useful goal as the sample set is balanced enough among emotion categories to train a NN to control a variety of NPC facial elicitations.

3 METHODOLOGY

The controversy over using actors to create samples or to collect samples of "real" subjects from streaming video websites has evoked important questions on the ethics and authenticity of emotion observation. But rarely in the discussion about the use of actors was there an in-depth discussion on the theoretical differences in beliefs about emotion generation between actors and psychologists, and among actors themselves. As performance requires communication between actors and those that "stage" their performance, beliefs about how emotions are generated will be implicit in the instructions given during preparation and rehearsal and in the method used to record the video samples.

3.1 Acting Theory Differences

There are some disadvantages to using actors for emotion elicitation video corpora that have been scarcely addressed by the communities that develop emotion corpora. 1) Actors are trained to communicate within a stylistic tradition of specific global regions, and those traditions, even those that strive for "realism", are not in fact inter-culturally "real"; they instead provide for the viewer of each region a broad set of compressed and coded elicitations that suggest emotional ideas for efficient storytelling. 2) The characteristics of an actor's performance is partially shaped by a director who is trained to design rehearsal procedures to elicit desired affects. It is broadly agreed among contemporary acting teachers and theorists that actors cannot simultaneously have acute awareness of their own emotive behaviour while also concentrating on the fictional stimulus that triggers it (Barr, Kline and Asner, 1997). The director is therefore relied on to see the performance in progress and to request adjustments from actors. Thus, any emotion elicitation experiment depends on a director to setup and modulate the actors. Therefore, to some extent control of the performers is not fully in the hands of the psychological and computational investigators who design the corpora. 3) Actors choose to be trained in one or more of a variety of methods or styles that require "belief" by the actor in the technique they have chosen to learn (Barr, Kline and Asner, 1997). These differences between techniques can complement or clash with each other, and conflicts can occur between actors or with a director who does not share the same "belief" or is unfamiliar with a particular approach that intends to achieve the illusion of authentic emotion elicitation in a fictional context. 4) Some of the underlying intentions of the discipline of computational psychology contradict some of the beliefs of major acting theorist that infuse much of the training of actors in academies and universities around the world. Specifically, the idea that an elicitation from an actor to make a video sample for a corpus need only be a demonstration of an elicitation, but not a demonstration of actual felt emotion. This conflict is widely known in the acting theory circles as the debate between "outside-in" or "inside-out" training methods. An investigator positioned on the opposite side of their actor's beliefs

might be unable to contribute to a corpus production project.

Despite these potential challenges, and in the light of the relative rise of enthusiasm in the affective computing and computational psychology communities for "in-the-wild" corpora, GEMEP, IEMOCAP and CreativeIT were created using actors with scripted and improvisational scenarios that target specific emotions for the purpose of creating comprehensive baseline definitions for emotion identification, close analysis of the processes of emotion generation, and inter-agent emotion interpretation. The results of the work of these corpora demonstrate that emotions targeted by researchers through experimental design, and then elicited by actors sufficiently prepared, can be identified and cross-validated by human evaluators. It is the intention of this research to implement the most effective and relevant methods of the creation of the GEMAP, IEMOCAP and CreativeIT corpora, and to augment their methods with additional parameters that better suit the needs of training of neural network modules that animate the face of game characters. Previous work with NN-driven animation have already shown that a small and rough sample of "invented" player emotion data for four basic emotions can drive the facial emotion animation of an NPC (Kozasa et. al 2006), or team of NNs can be trained to animate each specific emotion (Schiffer, 2021). One of the goals of this research was for the result to provide enough expressive nuance that the facial movement could be comparable to the performative aesthetics of live-action film and television. To accomplish this, several corpora of video samples were produced using professional actors directed by an experienced director familiar with the performance preparation methods that the collaborating actors had previously trained.

3.2 Active Analysis and the Repetition Exercise

Asking an actor to directly elicit an emotion, with or without a discussion of causality and context is widely understood among actors as results-oriented directing and is usually disparaged as cause to an "unnatural" performance (Weston, 2021). Corpus validation for acted elicitation has shown success when actors were allowed to prepare in the manner they were trained (Douglas-Cowie, Campbell, Cowie and Roach, 2003 and Bänziger, Mortillaro, Scherer, 2011).

For trained actors and directors of performers, emotional elicitations are deemed "natural" and "authentic" when they are the result of an actor fully believing the conflictive facts of the past, sensations of the moment, and projections of the future of a character. "Natural" and "authentic" performances occur when actors use beliefs to pursue character goals and circumvent obstacles while using the physical and emotional properties of the character. Inventing the past, present, and future of a character is a collaborative process between director and actor. But only the director, through their interpretation of the written script, can coordinate all the pasts, presents and futures of all the characters in a scenario to ensure that as characters pursue goals and circumvent obstacles, appropriate character conflicts arise for the actors with little effort. By guiding characters toward pursuing conflict-generating goals, an actor's valence (pleasure and displeasure) and arousal will elicit a full range of emotion elicitations so long as they are listening intentionally to the fictional sensations of the scene portrayed. Creating an elicitation without the coordinated context that a director is obliged to setup through imaginative blocking and explanation, deprives the actor of arousing sensations (real or imagined) from which to respond authentically, and can cause an actor to project "prototypical" emotion elicitations that appear unnatural or inappropriate to the scene's context. Such elicitation on demand, or result-direction, was avoided in the experiments of this paper. Instead, as studies conducted by Scherer, Bänziger, Busso, and Metallinou each propose, using preparation notes, discussion, and rehearsals to provide an actor information about the character, proved more effective for creating imaginary fictional stimuli. Such information and stimuli allow the actor to imagine the impetus of elicitation. Resulting elicitations and their annotated categories have been validated as mostly the same as their pre-selected target emotion (Bänziger, Montillaro, Scherer, 2011).

Of the acting theories explicitly referenced, as in Bänziger et. al and Metallinou et. al, Stanislavsky's two techniques of imaginative access of affective memory and his active analysis were both used in the creation of the corpus created for this paper. Actors were provided a scenario with a past backstory of facts, a current situation and three future outcomes that have vividly different imaginary conditions to provoke distinct emotion categories and diverging valence and arousal responses. Past events in the actors' histories that were similar in remembered sensation and emotion were tapped as affective memories.

Preparation of the actors included providing the basic parameters of active analysis for the actor

playing the role of the on-screen non-player character and for the game player character. Those parameters evolved through what Stanislavsky calls improvisational etudes, or repeated sequences (Thomas, 2016), to discover, 1) goal selection and acquisition through action with the use of tactile noun goals and physical verb actions, 2) possible tactical actions for circumventing obstacles and exploiting aids, and 3) development of imagined sensations that construct the fictional appraisal of the environment and a determination if actions previously taken are leading toward goal acquisition.

The experiment consisted of a dyadic conversation with restricted physical limitations and pre-defined social relations between the on-camera actor and an off-camera actor. As the scenario's scripted action and dialog progressed in time, obstacles, and aids to the characters in the scenario increased and resolved the conflict. Improvisation of all the non-lexical features of the scenario guided the actor toward a clear direction for every recorded sample. Figure 1 presents one of the actors in-character during production of a single-actor corpora in a studio environment with timecode slate. Figure 2 illustrates the setup of the recording of each sample.

Figure 1: Corpus Actor Alfonso Mann.

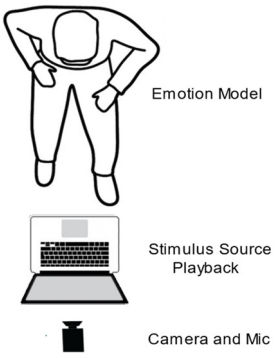

Figure 2: Production setup for sample recording.

While neither of the publications that describe the IEMOCAP nor MSP-IMPROV corpora explicitly mention Meisner's techniques, their descriptions of their own corpus production method and its emphasis on variation and repetition resemble Meisner's repetition exercise (Busso et. al 2008, Busso et. al 2017 and Meisner, 1987). The experiment of this paper deployed the technique by designing a dialog-behaviour tree where the lexical content is fixed but is reperformed several times such that some of the backstory facts of the character changed, the character goal for the scenario changed, the actions to acquire the goal changed, and the set of likely future outcomes changed as well. The dialog-behaviour tree shown in Figure 3 illustrates an 8-node directed acyclic graph with one distinct start node and one distinct end node. Each node represents a single or pair of dialog turns between the two characters in the dyadic conversation. Each edge represents the internal emotional progression the actor takes that leads up to the dialog turn.

The dialog-behaviour graph works as follows. The grey box segments represent the dialog of an Interrogator character henceforth called the Stimulus Source and the white box segments represent the dialog of a suspect character that we will refer to as the Emotion Model. The Stimulus Source starts at node 0 with the first turn. The Emotion Model either responds with a "Yes" down the Cooperate Path as shown in node 1 or a "No" down the Resist Path as seen in node 2. If the Emotion Model chooses to repeat the previous response of either "Yes" or "No," then the scenario advances directly down an edge remaining on the same path. If the Emotion Model decides to answer with the opposite response of either the Cooperate Path ("Yes") or Resist Path ("No"), then the Emotion Model switches across to the opposite parallel path down a diagonal line that advances to the next level of dialog turns. Another possibility is that the Emotion Model can respond with an intermediate "So". In this case, the Emotion Model moves across to the parallel opposing path but remains on the same dialog turn level. As this is a directed acyclic graph with a finite number of paths and path lengths, the Emotion Model cannot move to the next node on a previously traversed edge. The Stimulus Source terminates the graph with the final dialog turn at node 7.

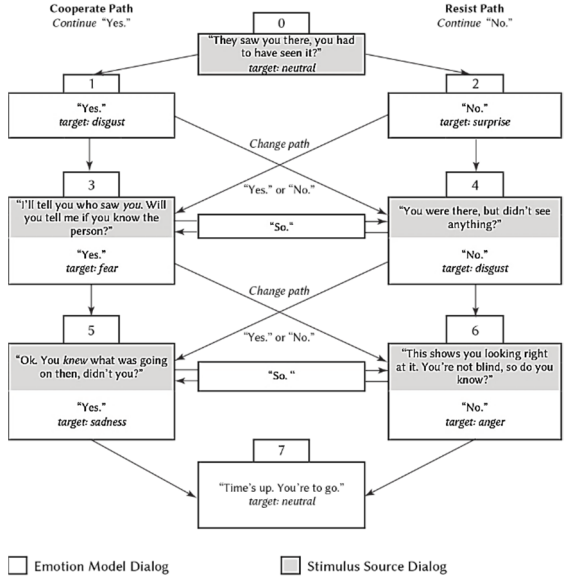

Figure 3: The 8-node directed acyclic graph produced 32 unique video clip variations. No nodes or edges repeat. Grey dialog was performed by the Stimulus Source. White dialog was performed by the Emotion Model.

Using a Depth-First-Search algorithm (Sedgewick and Wayne, 2011), we discover there are 32 possible paths without repeating any edges. For each path, the actors recorded 9 samples in 3 triplets. Thus, the total sample count was 288. The first triplet consisted of a version of the backstory and possible outcomes that would not likely disrupt the life of the character. The basic material and social needs that the environment contained would remain accessible and plentiful. This first triplet intends to provide a baseline of emotional normalcy with arousal scores seeking zero (mild boredom) and valence scores intending toward positive scores (pleasurable). This group of samples is the *low intensity outcome triplet*. The second triplet consisted of an embarrassing but legally unprovocative backstory with possible outcomes that could lead to short-term social exclusion, but the most negative outcome would have little effect on the future material state of the character, nor would it risk harm to their body. This set of samples was called the *medium intensity triplet*. The third triplet of video shots consisted of an illicit and morally depraved backstory with probable outcomes that could lead to extreme long-term social exclusion, loss of material possessions and personal freedom, as well as potential injury to the body. This third group of samples was the *high intensity triplet*. Despite each of these distinct scenarios, the lexical content was either totally fixed or slightly adjusted for each recorded sample. The differences in facts, actions, goals, and visualized outcomes provide impetus for variations in emotional elicitation.

To achieve consistency among each of the triplets, the performance director of the experiment adapted Clurman's script scoring technique (Clurman, 1972) such that for each dialog or behavioural "beat" (which in this case is represented as a node), there is one tactile goal represented by a noun (an objective), and one behaviour to get the goal represented by a verb (an action). Throughout the scene there is one perceived or imagined tactile obstacle and aid, represented by nouns. And lastly, there is a set of imagined outcomes where at least one is desired, leading to a positive outcome, and one feared, leading to a negative outcome. While the basic scenario for each actor playing the NPC was the same, the details that differentiated the triplets were adapted to fit to their physical characteristics, such as age, gender, physique, sexual orientation, and ethnicity.

3.3 Targeting Specific Emotions

Through their input device and logical controls, video games provide the player many opportunities to interact with the medium and directly affect the audio-visual stream of animation, sound, character behaviour, and thus narrative generation. Despite an abundance of aleatory features available to the medium, the player still looks for some pseudo-psychological cohesiveness to the behaviour of NPCs. Thus, when players observe an NPC's behaviour pattern in the game play, they expect variation to reflect the dynamic environment of the characters within the constraints of the game rules, including rules that govern social behaviours. Rapidly players notice behaviour patterns in NPCs that lead players to mentally generate narrative through cause-and-effect suppositions. Game designers exploit the tendency of players to mentally generate narrative by using the player's gaze in a close angle of NPCs' faces. Facial animation represents the elicited emotion of an NPC, and the emotion reveals by implication their thoughts and feelings about events, objects, and other characters, including the player.

By targeting specific emotions in the actor, a NN will train the weights and biases of its prediction algorithm to respond within the range of the values presented during the training stage. If the emotions generated in the scenario of the corpus production are within the desired range as those designed for game play, then the NN will predict and generate emotion data for the NPC that is appropriate to the game's narrative. To achieve this correlation, a dataset must be enriched by moderate variations of facial emotion

data of the same emotion category. This correlation was easily achieved in the design of the scenario by using a scripted and rehearsed dialog-behaviour tree.

3.4 Design of Emotion Elicitation Content

An examination of the structure of the dialog-behaviour graph reveals that, like any tree graph, it has a height and width dimension. In the use presented for the experiment of this paper, the height in stacked nodes represents levels of dialog exchanges, and the width of a layer of nodes along with edges that connects them, effects the number of possible choices a character could make through any path from the start to the terminus. As the graph grows in height, the average length of each video segment increases. As the tree expands in width, the number of possible variations of node-edge sequences increases, thus the number segments will increase. The rules of a directed acyclic graph forbidding repetition of edge traverses prevent reversal of progress from start to end nodes. This rule reinforces a principle of requiring a novel experience at every step toward resolution. Additionally, as the dialog progresses downward, the action by the Stimulus Source to acquire their goal becomes more aggressive. This aggression can cause the Emotion Model to react with an impulsive "Yes" or "No", or if the Stimulus Source eases pressure, the Emotion Model response can thoughtfully reflect before speaking.

At the first level, the Stimulus Source begins with a statement of facts and a suggestion of implication, "You had to have seen it." At the second level, the Stimulus Source offers a cooperative deal with "I'll tell you… Will you tell me?" or with an escalation of the surprise that the Emotion Model "…didn't see anything." The third level is the climax where the Stimulus Source accuses the Emotion Model of knowing but withholding testimony of the facts of a crime with either the more affirmative, "You knew…" or with the aggressive demonstration of evidence that ties the accused Emotion Model to a past crime scene. The final level is a plausible resolution where the allowed time of interrogation expires. This shift allows the Emotion Model to reflect on the events that just occurred and eventually return to neutral. The return to neutral must be accompanied by the cease of stimuli and then the return of the FER system to reset back to zero to match the starting point. Matching start and end node emotions are important for NN training as all the FER-produced data may be concatenated to a large flat-file database. Starting and returning to zero eliminates large value discrepancies between adjacent rows of data. Large data value jumps between adjacent rows can reduce a NNs trainability.

3.5 Sample Recording and Data Generation

The OMG-Emotion dataset emphasized the importance of corpus sample production methods that provide contextual data around facial elicitation (Barros et. al, 2018). Context comes in two forms: a representation in the data of what occurs before or after the target elicitation, or a representation that presents itself in the synchronous proximity of the target elicitation. For video samples, this concept of context would seem to refer to events that occur in previous or subsequent frames around the region of frames where the elicitation occurs. Or it refers to events, objects or agents that appear in the frame with the elicitor.

There is another kind of context though. There are events, objects and agents that occur or coexist synchronously out of frame but in the near proximity of the elicitor. For a dyadic conversation sample recorded for a facial emotion recognition corpus, the camera must be placed directly in front of the elicitor to optimize facial surface lines of sight into the lens and to minimize occlusion and data loss. But with one camera squarely in front of the eliciting Emotion Model, the opposing Stimulus Source is the most relevant item in the context of the elicitation. This presented an important question in the design of the sample recording procedure of the experiment of this paper. If the experiment will use the data of the elicitation context, should the Stimulus Source also be recorded with a second synchronized camera for every sample recording? Or should the Stimulus Source be pre-recorded such that all 32 paths through the dialog behaviour graph are presented during the performance of the Emotion Model on a video screen with speakers? Since the Stimulus Source is playing as the player character, the input data of the player must narrow down to a small enough set of predictable values so that the player can make either a reasoned or impulsive decision. Thus, the set of 32 paths would provide constant values frame-by-frame against all frames of the three triplets of recording samples of the Emotion Model. By keeping constant the Stimulus Source through the playback of a pre-recorded performance, the context of the elicitation is provided for synchronous annotation and analysis by a FER system. Additionally, the Stimulus source data is also available for training data for the NN model of

the NPC. Because the Stimulus Source speaks and gestures to the Emotion Model through a video screen, the Emotion Model modulates arousal and valence in response and in synch with each turn of the Stimulus Source. For each of the 9 variations in the 3 triplets of emotion recognition frame instances of the Emotion Model, there is 1 constant causal frame corresponding from the Stimulus Source.

4 RESULTS

Unlike the corpora reviewed for this paper, no human evaluators were employed to annotate the video samples. The reason for this difference is that a single-actor for single-character corpus is not designed to study emotions, nor is it developed to create emotion recognition applications. The purpose of annotating a corpus to be used for training a NN is to provide classification definitions of emotion categories of an actor playing the NPC so that the patterns found in a video game player whose face is not the Stimulus Source, but is outside the training, validation, and testing set, is correctly classified, and then appropriately responded to by the NPC facial mesh during the game. Quite different from classification corpora, a single-actor corpus is a simulation corpus. A simulation corpus depends on classification corpora to classify the context and predict the elicitation of a specific face that will be simulated. Thus, to cross-validate with human annotators would simply be a check on the FER system that classified the single-actor and would reveal little about the validity of the single-actor corpus.

There is however a way to validate the corpus. Targeting emotions in the design of the scenario will likely produce a corpus that will more easily train a NN model that can control the facial animation of an NPC. The FER-generated data must be post-processed into a data query that captures all sample segments that commonly cross the same edge. For each edge, there is a set of corresponding video segments bounded by the same timecode in and out points indicating the exact first and last frame the FER analysed for each of the segments. A presentation of data from all the edge segments of this experiment would be unnecessary and require more space than needed to demonstrate the efficacy of the method proposed. Instead, one typical edge segment shown in Figure 4 is edge D. Figure 4 shows a node-to-edge version of the dialog-behaviour graph. Edge D is among a set of edges selected from the experiment to demonstrate how to validate the set of FER data in relation to the target emotion intended for the edge.

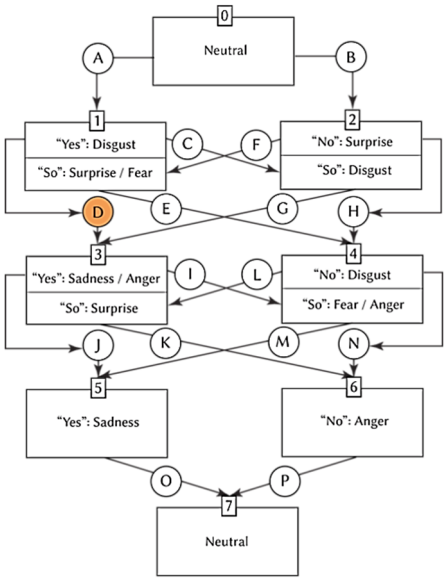

Figure 4: Dialog-behaviour graph for dyadic interaction. Edges (A to P), nodes (0 to 7). Model responses in quotes. Orange node-D is the focus of this study.

Table 1 shows the result of a sample data query where the enumerated paths appear in the row at the top of the table as column labels. The column numbers in Table 1 represent path IDs (1-32). Beneath the path ID is a set of numbers that show the order of numbered nodes in each path (0 to 7). Notice that each path that crosses edge D includes the consecutive presence of nodes 1 and 3. As shown in Figure 4, edge D is formed from nodes 1 and 3. Observe that paths 5, 6, 9, 10, 21, 22, 25 and 26 all traverse edge D. The ellipses (…) indicate omitted paths that do not traverse edge D.

Table 1: Instances of Edge D in Each Sequence.

		…	Seq. 5	Seq. 6	…	Seq. 9	Seq. 10
Nodes	Timecode		013457	01357		0213457	021357
1	0:14.6	…	1	1	…	1	1
3	0:26.6	…	1	1	…	1	1

…	Seq. 21	Seq. 22	…	Seq. 25	Seq. 26	…	Sum of Instances
…	013467	0136		0213467	021367	…	
	1	1	…	1	1	…	8
	1	1	…	1	1	…	8

From the data sort, the next step is to segment and concatenate the query result row data from the master data frame that only contains emotion values for the

timecode that covers edge D. Since all Emotion Model samples were recorded precisely in response to the same frames of the same Stimulus Source recording, the video frames of each Emotion Model response are precisely aligned in time-series.

4.1 Analysis and Validation Method

With FER data of the Emotion Model's response to the same stimulus during the edge, statistical results can validate the degree to which the Emotion Model's performance met edge D's target emotion, as Figures 5 and 6 show. Statistical analysis over all the values produced by frames analysed by the FER system in the edge of each path in the dialog-behaviour graph can indicate if the target emotion elicitation was adequately collected by following two steps: 1) Find the proportion of values above a primary threshold among all samples that traverse a given edge, 2) Find the mean of values at each frame among all the samples that traverse a given edge, organized by categories of low, medium and high intensity in triplets of positive to negative consequence.

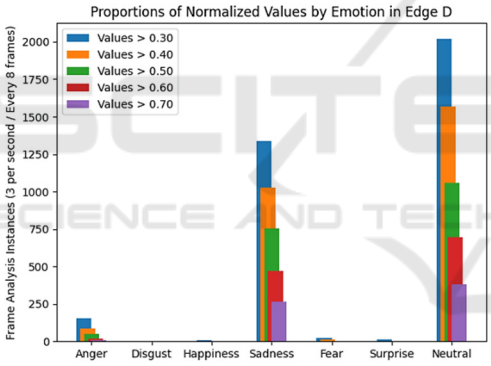

Figure 5: Performance of edge D reveals sadness is primary emotion, followed by anger.

Figure 4 identifies the position of edge D in the dialog-behaviour graph as an edge leading into node 3. The primary emotion is sadness, and the secondary emotion is anger. Figure 5 reveals the proportions of normalized values in edge D leading into node 3. The disproportionate instances of neutral emotions are a common feature of facial elicitation emotion value histograms. Listening and thinking often show high neutral measurements. Figures 6 and 7 show proportions of emotions *sadness* and *anger* over the same segment of timecode.

Figure 8 confirms the higher levels of *sadness* lingering with a gradual decline that gives way to *anger* for a climactic reaction as shown in Figure 9.

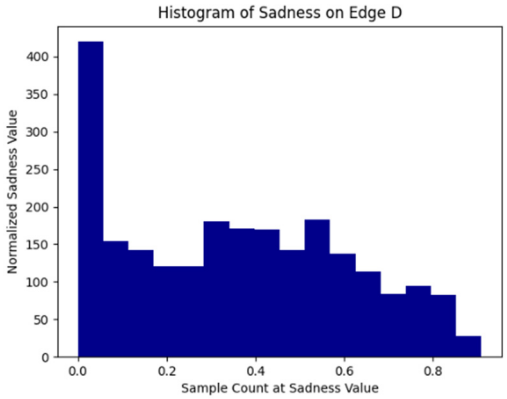

Figure 6: Histogram of primary emotion sadness for edge D indicates maximum is 0.910 and the median is .384.

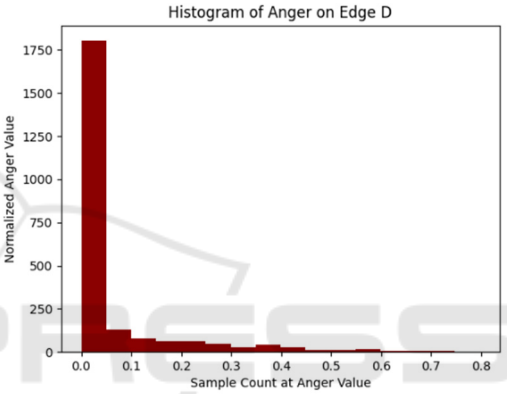

Figure 7: Histogram of secondary emotion anger for edge D indicates maximum is 0.797 and the median is 0.0000006.

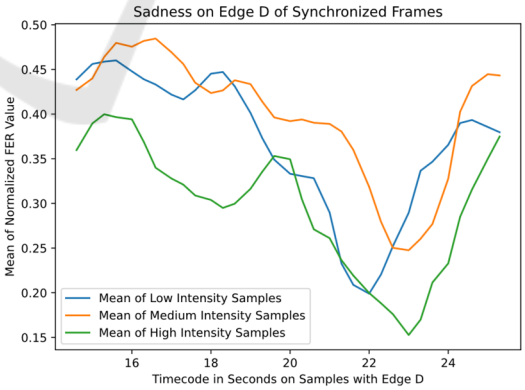

Figure 8: Primary emotion *sadness* on edge D wavers and rises, giving space for secondary emotion.

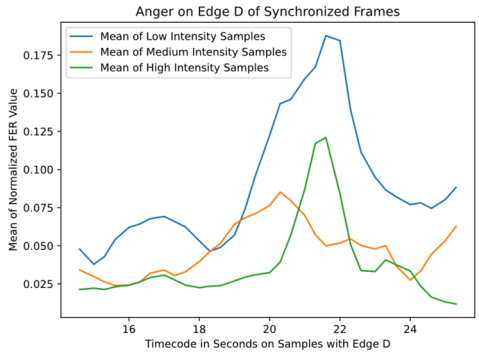

Figure 9: Secondary emotion *anger* on edge D asserts peak.

It should be noted that the intention of the intensity triplets was not adhered. The Low Intensity Samples showed the highest values, Medium plots the lowest and High falls in between. Production conditions sometimes yield such deviations. The dynamic range and higher maximum value of sadness indicates that it is the dominant emotion of the video segment, thus validating that the target emotion indicated in Figure 4 for edge D was elicited in the segment intended.

5 DISCUSSION

This paper disclosed at its Introduction that neural network design details and training techniques are outside its scope, but their design properties and their training methods inform the characteristics of an optimal single-actor corpus. An optimal corpus will allow a NN to train without over- or underfitting.

Designing the dialog-behaviour tree with the intention of producing emotional variability can prevent overfitting. The dialog-behaviour tree will determine the variability of emotion values generated from the Emotion Model. Most important is the variance of values for each emotion category (*anger*, *sadness*, *fear*, *disgust*, etc.). By providing sufficient branching variations in the dialog-behaviour tree, unsupervised learning can train a NN without overfitting. But will optimization best be served by adjusting the only the width of the dialog-behaviour graph? Or will longer samples with more time-series or nodal steps, thus increasing a tree's height, provide data that is variable enough over time to avoid under-fitting? Or can expansion of height and width of the dialog-behaviour tree in some proportion improve performance of the NN? Each of these parameters will hint at the benefits or detriments of longer samples or more varied samples.

When producing facial emotion corpora for this research, some FER biases were observed. However, all FERs on which initial emotion analysis is based, are subject to biases of reading facial characteristics of the actor. Techniques for countering FER bias is a topic for another investigation, but it must be considered in designing facial emotion corpora. These biases can also be amplified by uncontrolled visual conditions, such as poor lighting or optical distortions from wide angle lenses, another obstacle for future research to consider. Lastly, as subsequent investigations analyse which FERs are most accurate, one should recognize that FERs frequently use both convolutional neural networks (CNNs) for classification of static object data, such as individual frames of faces and the features of the face that do not change (e.g., bone structure). More recently FERs use recurrent neural networks (RNNs) for time-series prediction of facial features that are dynamic (e.g. movement of muscles and flesh). Ideally, training a NN with FER data that uses CNN and RNN components, given their distinct architectures, may allow time series data to balance the emotion readings of face-structure with those of facial motion.

6 CONCLUSION

Video game character animation has long been the domain of animators who in the past scarcely used actors for character design but now use them frequently for motion and volumetric data capture. Motion capture has expedited and enriched photorealistic video game design with cinematic performance aesthetics for its complex facial emotion expression. FER systems can also contribute to this field of video game development through the development single-actor corpora that capture intangible and surprising patterns in detailed human expressivity controlled by emotion AI. The production principles of single-actor corpus design for the development of NN models for NPC facial expression animation are currently in a nascent state. Much remains to be discovered on how to design the sample production method to precisely acquire the desired results. Open questions remain concerning the variations of preparing the role with the actor for emotionally or narratively complex video game scenarios and the NPCs that populate them. Lastly, a comprehensive comparison between current costs of NPC animation and projected costs for a combined corpus and neural network solution remains to be completed before any declaration that this novel method is in fact viable for commercial application.

REFERENCES

Bänziger, T., Mortillaro, M., and 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.

Barr, T., Kline, E. S., and Asner, E. (1997) *Acting for the Camera*. Harper Perennial, New York, NY, USA, 8-18.

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). doi: 10.1109/CGIV.2006.41.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S., (2008) IEMOCAP: Interactive emotional dyadic motion capture database. In *Language Resources and Evaluation*, vol. 42, no. 335. https://doi.org/10.1007/s10579-008-9076-6

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi. N., Provost, E. M. (2017) MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. In *Transactions on Affective Computing*, vol. 8, no. 1. Jan-March 2017. IEEE, New York, NY, USA. 67-80.

Clark, J. (1996) Contributions of Inhibitory Mechanisms to Unified Theory in Neuroscience and Psychology. *Brain and Cognition*, vol. 30, 127-152.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P., (2003) Emotional speech: Towards a new generation of databases. In *Speech Communication*. Vol. 40. Elsevier, New York, NY, USA. 36.

Feldman Barrett, L. (2017) The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, vol. 12, iss. 1, January 2017.

Clurman, H. (1972) *On Directing*. Fireside, New York, NY, USA. 80.

Kozasa, C., Hiromiche. F., Notsu, H., Okada, Y., Niijima, K., (2006) Facial animation using emotional model. In *Proceedings of the International Conference on Computer Graphics, Imaging and Visualization* (CGIV'06). 428-43.

Meisner, S., (1987) *Sanford Meisner on Acting*. Vintage Random House, New York, NY, USA. 26-38.

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 Corpora* (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. 41-46.

Scherer, K. R. and Bänziger, T. (2018) On the use of actor portrayals in research on emotional expression. In *Blueprint for Affective Computing: A Sourcebook*, eds. K. R. Scherer, T. Banziger, and E. B. Roesch Oxford Univ. Press, Oxford, England. 166–176.

Scherer, K. R. (2013) Vocal markers of emotion: Comparing induction and acting elicitation. In *Computer, Speech Language*, vol. 27, no. 1, Jan. 2013. 40–58.

Schiffer, S. (2021) Game Character Facial Animation Using Actor Video Corpus and Recurrent Neural Networks. In *International Conference on Machine Learning Applications* (ICMLA 2021), December 13-16, 2021.

Sedgewick, R., and Wayne, K., (2011) *Algorithms*, 4th Edition. Addison-Wesley, New York, NY, USA. 570-596

Strasberg, L. (2010) Ed. Cohen, L., *The Lee Strasberg Notes*. Routledge, New York, NY, USA. 47, 149.

Thomas, J. (2016) *A Director's Guide to Stanislavsky's Active Analysis*, Bloomsbury, New York, NY, USA.

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.

Weston, J. (2021) *Directing Actors: Creating Memorable Performances for Film and Television*. Michael Wiese Productions, Studio City, CA, USA, 1-11.