5-1-2023

An Actor-Centric Approach to Facial Animation Control by Neural Networks For Non-Player Characters in Video Games

Sheldon Schiffer

An Actor-Centric Approach to Facial Animation Control by Neural Networks For Non-Player Characters in Video Games

by

Sheldon Schiffer

Under the Direction of Ying Zhu, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Game developers increasingly consider the degree to which character animation emulates facial expressions found in cinema. Employing animators and actors to produce cinematic facial animation by mixing motion capture and hand-crafted animation is labor intensive and therefore expensive. Emotion corpora and neural network controllers have shown promise toward developing autonomous animation that does not rely on motion capture. Previous research and practice in disciplines of Computer Science, Psychology and the Performing Arts have provided frameworks on which to build a workflow toward creating an emotion AI system that can animate the facial mesh of a 3d non-player character deploying a combination of related theories and methods. However, past investigations and their resulting production methods largely ignore the emotion generation systems that have evolved in the performing arts for more than a century. We find very little research that embraces the intellectual process of trained actors as complex collaborators from which to understand and model the training of a neural network for character animation. This investigation demonstrates a workflow design that integrates knowledge from the performing arts and the affective branches of the social and biological sciences. Our workflow begins at the stage of developing and annotating a fictional scenario with actors, to producing a video emotion corpus, to designing training and validating a neural network, to analyzing the emotion data annotation of the corpus and neural network, and finally to determining resemblant behavior of its autonomous animation control of a 3d character facial mesh. The resulting workflow includes a method for the development of a neural network architecture whose initial efficacy as a facial emotion expression simulator has been tested and validated as substantially resemblant to the character behavior developed by a human actor.

INDEX WORDS: Emotion AI, Non-player characters, Video games, Neural network animation controller, Autonomous animation, Facial emotion video corpora, Facial emotion recognition

An Actor-Centric Approach to Facial Animation Control by Neural Network

For Non-Player Characters in Video Games

by

Sheldon Elias Schiffer

Committee Chair: Ying Zhu

Committee: Rajshekar Sunderraman

Zhisheng Yan

Gregory Smith

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2023

DEDICATION

I am eternally inspired by my daughter, Ariella Miranda Schiffer, whose birth at the beginning of this research project, and whose survival and thriving throughout is an experiment of far greater importance than anything I could ever do. I am humbled by Nelly Fuentes, my partner and spouse, who has endured our long haul of making this dissertation during our first decade of knowing each other. I must share that my mother Gloria Arianna Schiffer, has been a great cheerleader and guardian, for whom I am grateful for protecting me from distractions and providing me notices of the hazards of neglect of the basics-of-life that intense study can tempt. For all that they have given to make this research possible, I dedicate this work to my family, for whom I work to make this project a success, and who reminds me why we feel emotions and desire to share them with the world.

ACKNOWLEDGEMENTS

Among my professional colleagues, I must acknowledge the support and guidance of my advisor, Dr. Ying Zhu who has been immensely helpful at helping me fine tune the writing that describes complicated things, and at finding other colleagues out in the world who may care to read this work. Dr. Zhu introduced me to Max Levine and Samantha Zhang, both National Science Foundation mentees who helped model and code the non-player characters used for animation in this study and were co-authors of one of the chapters. I am also grateful to Dr. David Cheshier and Brennen Dicker, both who encouraged me and supported my continued study while overseeing our common dedication to the burgeoning Creative Media Industries Institute at Georgia State University. I am grateful to all of the faculty of the Department of Computer Science who welcomed artist-become-scientist into their classrooms and offices. I must also express my thanks to the actors Alfonso Mann, Marilyn Sanabria, Amir Kovacs and Keith Tims who helped me create the facial emotion corpora for this study. Their willingness to participate, albeit cautiously during a pandemic, made all the difference.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADFES	……	Amsterdam Dynamic Facial Expression Set
AU	……	Action Units
BD-LSTM	……	Bi-Directional Long Short-Term Memory
FACS	……	Facial Action Coding System
FER	……	Facial Emotion Recognition
GEMEP	……	Geneva Multimodal Expression Corpus
IEMOCAP	……	Interactive Emotional Dyadic Motion Capture (database)
LST	……	Long Short-Term Memory
ML	……	Machine Learning
MSP-IMPROV	……	Multimodal Signal Processing Improvisation (database)
NN	……	Neural Network
NPC	……	Non-Player Character
OCC	……	Ortony, Clore & Collins
RNN	……	Recurrent Neural Network
SBFSM	……	Stack-Based Finite State Machine
WSEFP	……	Warsaw Set of Emotional Facial Expressions

LIST OF LISTINGS

1 INTRODUCTION

The reaction among some in the arts communities to autonomous machine-derived creative production has ranged from skepticism toward its aesthetic value, to hostility toward its ambiguous and potentially intrusive role in the workflow of artistic production. Much of the skepticism and hostility has been a response to implementations of AI that segregate artists away from the design of the sampling process for machine-learning, and instead positions artists as only producers of content whose patterns of expression can be freely sampled by machine-learning algorithms with neither the artist's consent nor participation in the sample creation process. Granted, much of the artistic product used in the most current iterations of AI-derived artistic product has been from artists who are long deceased. But such a limiting and potentially exploitative relationship could only be a formula for resentment and decline of innovation. At the time of this writing, AI-derived creative products generated in the style of graphic or literary artists have been the primary subject of global fascination and acrimony. It is only a matter of time before the performing arts of music and acting become new material for autonomous machine emulation. Eventually, they will sample existing digital corpora of deceased or living performing artists to train neural networks (NNs) to play musical instruments or 3d modeled characters to appear and play as known actors or characters. The current model may exploit large samples of music or acting and eventually yield performances that satisfy audiences. AI in acting and music may also disturb audiences with the awareness of the machine-derived source of the artifice. The passive role that positions artists as an exploitable resource for technology is neither inevitable nor ethical. Artists can master the process of designing corpora as they have many other complex media, thus making the audio or video corpus itself a new medium of artistic expression intended for machine-learning algorithms that collaborate with an artist's intentions.

A starting strategy for this approach is found among the art forms and scientific practices that already have well-developed theoretical foundations that organize emotion analysis based on observing the physical and psychological patterns of human subjects and the actors who portray humans. Performance theory offers a method of training and preparation of the performer toward a tangible performance, while using iterative techniques of structured repetition. This research integrates a variety of techniques into a workflow for the development of autonomous facial animation. These include methods frequently used in the preparation an actor makes for a role in theater, film, and television and video games. For this study, it includes the creation of a corpus of video clips that will be used to train a NN that will animate a photorealistic 3d non-player character (NPC) model. The NN ultimately controls the animated facial expressions of an NPC performed by the actor in-character. Annotation as preparation for theatrical or screen performance often demonstrates how an actor uses their internal emotional behavior as an impetus to create character-specific movements, postures, gestures, and facial expressions – all embodied elicitations. Annotation is also used to determine the emotion labels and their intensities for recognition and simulation of emotion by training NNs for autonomous synthetic agents such as NPCs.

In addition to the performing arts disciplines, the field of psychology has also investigated and theorized extensively the human process of emotion generation. Fictional performance of facial emotion expressions communicates (albeit somewhat ambiguously) a character's "appraisal" of three perceivable categories of objects in terms that either affirm or interfere with a character's beliefs, desires, and intentions. Those objects are *events*, *agents*, and *objects*. These three components are the principal stimuli for emotion generation as defined by Appraisal Theory, a widely referred psychological model used to describe the feedback system

that generates emotions within humans. Appraisal Theory explains how emotions are generated in human interactions, in terms that describe emotion as observed phenomenon occurring in the course of “real life”, as opposed to the artifice of life as experienced in theater, film and television.

Appraisal Theory and Performance Theory refer to the same human system of emotion generation, but for distinct purposes. The centuries-old Performance Theory developed to train and guide actors creating fictional characters. Appraisal Theory evolved to both describe a system of human emotion generation, and to provide a blueprint for simulating human emotion in a computer. Inevitably, the two theories can be considered together for the creation of autonomously emotive humanoid characters, either in the form of a 3d model for a computer-driven interactive experience, such as a video game or conversational agent, or for a life-sized android machine in humanoid form. Psychology vis-à-vis Appraisal Theory provides a model for generating emotional *responsiveness* in synthetic agents – generating physical elicitations that appear to externalize the feelings we experience when something happens to humans. Theater and screen media vis-à-vis Performance Theory provide a model for emotional *agency* in artificial agents – generating the feelings humans experience when engaged in an action to get a goal. Responsiveness (from external entities) and agency (toward external entities) are thus two fundamental poles for elicitation that determine which template to refer to when annotating a script for production (planning agency), or when annotating a video clip within a corpus (documenting responsiveness).

Even as technology has progressed toward simulating human forms and movements using 3d polygonal meshes of higher polygon counts, simulating human facial expression movement with computationally enabled hand-crafted animation remains a standard artistic

practice through the late 20th and early 21st centuries. Referencing emotion annotation methods for 2d, and later 3d animation, was and remains a widespread practice that relies on late 19th century theories of elicitation and emotion signification developed for oratory and theatrical performance. The emergence of technologies that sample large video and static image datasets of human expression for NN driven animation, requires a distinct technique to preserve the authorial design of actors who are creating fictional characters for a new age and media.

The purpose of this research is to introduce a method for future animation development of NPCs in video games. We believe that video game NPCs offer the most appropriate medium for NN controlled facial animation trained on single-actor corpora. Current systems use generalized facial emotion elicitation that approximate what some researchers in cognitive psychology and its extension into media studies call “emotion prototypes.” But a generalized facial emotion corpus provides no path for the actor as a character author to provide elicitation idiosyncrasies that make for memorable expressions that audiences and game players distinctively remember. Our intention is to give the actor this path, and to foster a more collaborative workflow over a more iterative production process. With this approach, as a character matures with its story world through its actor’s behavioral adjustments, its corpus will also expand and provide more intricate patterns to train a NPCs NN facial controller.

As the chapters in this body of research will reveal, we show that a NN architecture can be designed to adapt to an actor’s role preparation and video corpus performance. We also show that two important measurable features can be used to demonstrate substantive facial emotion behavioral resemblance – *emotion intensity* and *emotion velocity* (defined as the rate of change of intensity over units of time measured in video frames). While our research reveals that both properties have been discussed as a means to validate NNs that recognize emotion elicitation, we

found very little research that uses them to simulate expressions in an animated NPC. In this respect, our research is unique and contributes to the growing field of affective computer applications within interactive entertainment.

We have been able to accomplish our research objective because of major technological advances during the previous decade that were not available when ideas of autonomous emotion elicitation first arose. The interest in developing autonomous emotion elicitation in synthetic agents has persisted and grown since at least the mid-1990s, but only current microprocessor technology has allowed for off-the-shelf computers to be used for relatively short NN training and validation sessions, and for the commercial development of emotion recognition systems of facial elicitation in real time video. Advances in specialized Python language libraries have made the use of annotation data from video corpora to be used as the dataset for training NNs to recognize emotion. All these aforementioned methods, among many others, we have taken into consideration to the service of developing a more actor-centric workflow for autonomous facial emotion animation. Our method was developed through the combined process of researching previous related experiments, and by designing and documenting the results of our own experiments. The reader may read the chapters in any order, though reading in the order provided here follows the workflow that was efficacious for our research goals. Before reading the chapters, it will serve the reader to familiarize themselves with the workflow diagram that appears in figure 1.1. Like most any computational process, ours can be introduced by summarizing the cultivated inputs, architected data structures and their algorithmic processes, and the anticipated data outputs. The workflow diagram gives a systematic description where each enumerated white development block can be affiliated with the chapters.

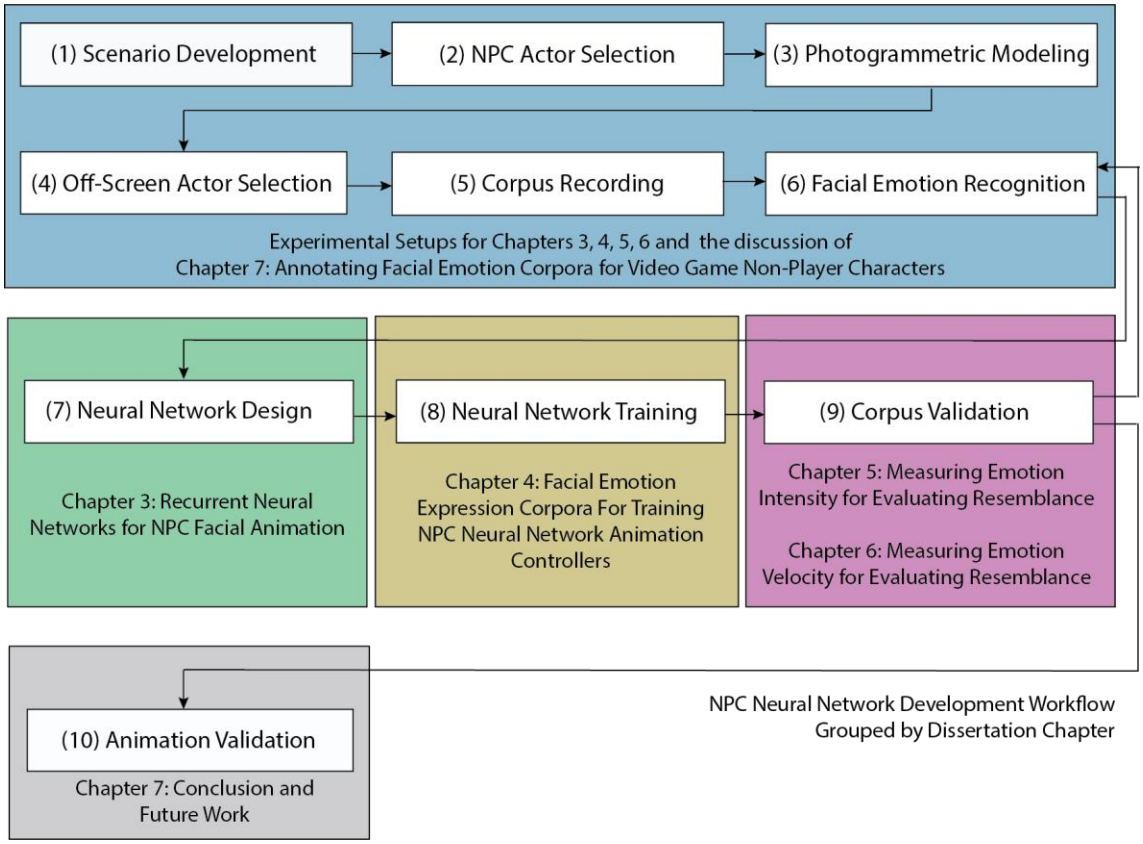

Figure 1.1: The Production Workflow Grouped by Chapters. Workflow for corpus production, neural network architecture resulting in interactive facial animation of an NPC.

For each of the chapters, we will summarize their relationship with the enumerated white development blocks of our workflow. They are described in each circumscribed chapter. (1) We develop a game play scenario between two characters as a dyadic conversation. The conversation is scripted as a dialog behavior tree in the shape of an acyclic graph. (2) We cast actors to play roles written in the script. The NPC role is also the Emotion Model for our experiments. (3) We deploy photogrammetric modeling by photographing the actor playing the NPC/Emotion Model and we construct a 3d head and face model with a photorealistic facial texture. (4) We cast an actor in the Player role. They interact with the NPC as the Stimulus Source for our experiments.

(5) We rehearse and record a corpus consisting of 288 clips with multiple takes for each path in an acyclic graph of a dialog behavior tree. (6) We conduct facial emotion recognition analysis at 3 frames per second for every clip, with video running at 24 frames per second, to create a dataset of emotion intensity values for six emotions and neutral. (7) We design a team of seven NNs, one for each emotion (happiness, sadness, anger, surprise, fear, and disgust) and neutral whose architecture is designed to optimize the character development process of actors. Chapter 3 describes our process of choosing a neural network architecture. We train, optimize, and validate the NNs through a process of comparing the various NN parameters and choosing the one with the least error. (8) We statistically validate the method of corpus production by closely examining NN performance on one highly typical segment of the dialog behavior tree. Chapter 4 elaborates on previous work in corpora production and the various methods used to produce and validate them. (9) We statistically validate the corpus by creating a standard of resemblance based on a comparison between predicted and observed data with a mean of means of frame-by-frame measurements of intensity and velocity for targeted emotions that occur in segments where those targeted emotions should appear. Chapters 5 and 6 demonstrate our method for determining statistical resemblance. (10) Animation validation is a step we propose for future research which we will discuss in Chapter 8 to conclude this discussion of our research. Chapter 7 concerns all chapters in this research. There we review the methods of creating data sets from corpora annotation used for NN training. We discuss the controversies that exist for this de-rigueur procedure. A more detailed summary of the chapters follows.

In Chapter 2, “Multi-Disciplinary Paths to Emotion Modeling: A Survey of Literature on Autonomous NPC Emotion,” we examine some of the earliest considerations of the problems of machine-generated emotion from philosophers, psychologists, storytellers and computer

scientists, and the most influential solutions that video game character designers have developed. The discussion brings the reader up to date with state-of-the-art concepts on computational emotion modeling for NPCs, and it also illuminates the methods we considered before developing our own experiments.

Our experiments in the domain of NN architecture are documented in Chapter 3, "Recurrent Neural Networks for NPC Facial Animation." This chapter details the most successful of many experiments in designing a NN trained on a single-actor corpus for the purpose of controlling a photorealistic facial mesh that closely resembles the actor.

Our NN architecture was optimized for a specific dataset whose production methods are elaborated in Chapter 4, "Facial Emotion Expression Corpora for Training NPC Neural Network Animation Controllers." This chapter details how our corpus production method evolved from many previous corpora we considered before designing our own approach.

While our first experiments with NN architecture and corpora production showed promising results, we needed a statistical method to validate our corpus and NN in a way that was distinct from standard NN validation. Typically, NNs are validated by a study of statistical difference between observed data instances and predicted data instances. Since our data set is time series data, basic methods of validation rely on ascending order analyses of error with the earliest instance of statistical analysis occurring first. But this basic approach would show no consideration for motivated and targeted emotion modulation in a scripted dyadic conversation tree performed through all conversation paths with many repeat performances from identical stimulus. Instead, we did a frame-precise synchronized error analysis among clip segments where the performing actor responded to the exact same synchronized stimulus. Stimulus of a predicted emotional characteristic and time instance were affiliated and compared with the NN

animation responses. What we describe in Chapter 5, “Measuring Emotion Intensity for Evaluating Resemblance” and Chapter 6, “Measuring Emotion Velocity for Evaluating Resemblance,” are two practical methods for performing statistical resemblance in emotion elicitation behavior to determine how closely a NNs facial animation control responds to the same stimulus as the actor did in one of the facial emotion corpora used to train the NN. By sorting data into partitions of performance segments from a dialog behavior tree shaped as an acyclic graph, we could cross-validate video segments where the performer responds to the same synchronized stimulus.

Chapter 7, “Annotating Facial Emotion Corpora for Video Game Non-Player Characters,” focuses on the annotation process that creates the datasets used to train the NNs. The entire enterprise of our research relies on a procedure of facial emotion annotation. It would be impractical and unfeasible to conduct our research and experiments using manual annotation as is commonly done with first generation general corpora used to develop NNs for FER systems. FER systems are created and revised with updated data sets in timeframes that are years long. The task of hand annotation with teams of human annotators is extremely laborious and therefore expensive. But, as a long-term research and development investment, such endeavors can provide a useful tool for facial emotion study for subsequent generations of scholars and scientists to use. One such use occurs in our research, which uses automatic annotation using a commercial FER system. But the underlying scientific premises of FER development are controversial. Within the fields of psychology and neuroscience, clear lines have been drawn that put Emotion Essentialist who believe in phylogenic emotion elicitation commonality among primates in opposition to Emotion Constructivist who believe that elicitation is not phylogenic, but that evolution provides a common set of neural structures for culture to manifest expression

types within the brain's elicitation and recognition systems. Our lasts chapter recognizes some of the dubious positions of the Emotion Essentialists, while also seizing on the notion that NPCs, like actors, are communicating authored fictional ideas in a structured manner designed to elaborate a narrative and are not themselves real people. Therefore, much like actors are trained to simulate and represent human characters in a manner that manifest narrative by communicating emotional states through the face, likewise NPC faces and the systems used to create data and train the NNs that control them, they need not present a "ground truth" emotion of a real human, but instead a much more circumscribed set of elicitation patterns that make up a "ground truth" of a character fabricated by an actor. This more limited ground truth is a dataset that is as subjective as any artistic invention one may engender about any human representation in an artistic medium. Our annotation method extends the process of actor-centric subjective character design from performance theory for actors, through the structured design of computational and cognitive psychology, into data structures useful in computer science and games research.

This research may also inform the computational social scientist who may design and study autonomously animated synthetic agents. The result will be a model for NPC facial animation that integrates methods used in the preparation of an actor's character and role in animation of any agent – the *agency* of emotion – with the methods used in the modeling of human emotion in the natural setting as analyzed by social and cognitive psychologists – the *responsiveness* of emotion. We hope the results will eventually lead to the continuation of deeply emotive cinematic acting in video games that is actor-centric, preserving a creative space for the actor as author for the future, and for synthetic agents in other media that use actors as a basis for NN-controlled facial animation.

2 MULTI-DISCIPLINARY PATHS TO EMOTION MODELING: A SURVEY OF LITERATURE ON AUTONOMOUS NPC EMOTION[1]

Video game NPCs are a type of agent that often inherits emotion models and functions from ancestor virtual agents. Few emotion models have been designed for NPCs explicitly, and therefore do not approach the expressive possibilities available to live-action performing actors nor hand-crafted animated characters. With distinct perspectives on emotion generation from multiple fields within narratology and computational cognitive psychology, the architecture of NPC emotion systems can reflect the theories and practices of performing artists. This chapter argues that the deployment of virtual agent emotion models applied to NPCs can constrain the performative aesthetic properties of NPCs. An actor-centric emotion model can accommodate creative processes for actors and may reveal what features emotion model architectures should have that are most useful for contemporary game production of photorealistic NPCs that achieve cinematic acting styles and robust narrative design.

2.1 Introduction

"Constructing characters" is a phrase that infers distinct meanings for two participants in the creative process of computer-based media. On one hand, for the narrative architect of video games or other kinds of narrative computational media, it is a semiotic process of fiction authoring where the character designer provides a written personal history. Three-dimensional or two-dimensional models of anthropomorphic shape combined with voice can also suggest agency in a game space. A weapon-wielding muscular humanoid with big bright eyes is well equipped for video game combat and all the emotional expression players associate with

[1] This chapter appears under a similar title. Schiffer, S. (2021). Multi-Disciplinary Paths to Actor-Centric Non-Player Character Emotion Models. In Bridging the Gap Between AI, Cognitive Science, and Narratology With Narrative Generation (pp. 17-42). IGI Global.

fighting. Once audio-visual elements are programmed to react to user input, these elements can signify to the observant player an imagined persona. On the other hand, for the developer of video game computer code, “constructing characters” is a process of designing a system that uses quantitative data derived from the game program, the computer operating system or from player input data to control or trigger character animation and voicing such that the player experiences the presence of a seemingly intelligent cohesive character identity. The result of the work of the designer and the developer is a composite signified that evolves in the player’s mind over the time of game play.

The manner of construction for narrative architects of video games depends on mental processes of player participation. Over the course of game play time, the player may observe actions and behaviors of NPCs so that a pre-game play biography and an in-game “alterbiography”[1] combines the NPC’s pre-game past with the evolving NPC actions the player witnesses or learns through game interaction since the start of the game. The manner of character construction for video game developers depends on computer languages whose frameworks contain data structures (primitives, classes, objects) and data behaviors (methods, functions) that can trigger and manipulate unique animations of the three-dimensional mesh model and its sound emanations (usually a voice) in ways that resemble the player’s understanding of human and animal emotional expressions. These NPC animations and sounds must be recognizable by the player as specific to the NPC’s type as a fictional narrative agent (human or non-human) and consistent with the character’s role within the world of the game.

The study of character construction coincides with the related research in other disciplines. The mid-twentieth century saw a confluence of ideas from disciplines that sought to provide taxonomies of two human endeavors – storytelling and emotion expression. Narratology

addressed the former and evolved from literary theory and folklore studies to describe systematically the human perception and representation of stories in various media. The categorization of characters within stories based on a typology of roles and emotion sets afforded to those roles is one specialization within narratology. Cognitive psychology and its subdiscipline, Computational Psychology, evolved as a reaction to Behaviorism and as an alternative explanation to the mental and emotional processes that drive human behavior. The categorization of human emotions as well as their neurological processes is one subdiscipline that cognitive psychologists frequently consider. The two disciplines converge in computer game design because game character designers and game code developers both use models from which characters can be efficiently produced. These characters and their behaviors can be embodied as preconfigured audio-visual animation and sound synthesis systems for “static” characters or can be used in-game to spawn procedurally generated characters or behaviors.

A discussion of NPC modeling benefits from the character theories of Narratology and the emotion theories of Computational Psychology because both examine the behavior of human or human-like agency. Narratology considers how foregrounding particular NPC behaviors in the context of an audio-visual story system forms the role of a character. These behaviors are a functional necessity for story development as they are elaborated in the player’s mind. Computational Psychology considers how to represent in computational form, a simulation of the internal processing machinery of the emotive part of the human mind so that when one implements and embeds an emotion model in an agent such as an NPC, the sensory input will yield an expressive output as an appropriate behavior that simulates a coherent human-like character. To design an NPC model of emotion for video games, one must consider how

emotions in characters are useful in the elaboration of narrative, and how they are generated by actors for use in animated characters.

Whether one considers video games as narratives or narrative systems, the fact a player *moves* through the time and space of a game world provides sufficient conditions to apply the frameworks of narratology. For games with NPCs, the player experiences over time the construction of character within their mind. Each interaction with an NPC is a momentary witnessing of in-game action with all its semiotic unwrapping over the duration of gameplay. Cognitive Psychology recognizes that expression of the body, and particularly those seen on the face, belies the mental properties and emotional states of a person[2]. It also communicates emotional meanings from the face of one person into the eyes of another. The latter function of emotion serves narrative expression in the visual and performing arts through the agency of character. The degree to which facial expression affects game narrative depends on the complexity of expression, and therefore the degree of immersion and believability a player may experience. Thus, an effective NPC emotion model must strive for complex expressions.

2.2 Narrative Expansion Through Facial Expression Complexity

Seymour Chatman gives a comparative study of Maupassant's short story *Une Partie de campagne* (*A Country Excursion*)[3] with Jean Renoir's rendition of the story in the film *A Day in the Country.*[4] Chatman provides persuasive analyses of how emotions within characters as described by words in a literary narrative are interpreted through cinematography and editing to convey similar ideas in a film adaptation.[5] Many of these emotional ideas that Chatman identifies depend on the framing of shots in the film where actors' faces communicate their emotional states. But cinema is not video game. In Renoir's film we are forced to see faces within a frame and in the order of a sequence of images. The dialectical process of montage allows the

meanings of faces to be refined and elaborated by the image that precedes a facial closeup and by the image that follows. In video games, only cutscenes lock the player into seeing specifically framed and ordered images of faces. When the camera is released to the volition of the player, there are few if any cuts to other images. Once the camera is free to move in response to the player's will, there is no way for the game designer to be sure that the player sees anything in a particular order. An NPC's facial expression must convey its ideas independent of other objects in the viewing screen. The need for the player to maneuver and see an NPC's face must be embedded in the mechanic of the gameplay. While lighting, music and voice can reinforce the emotional state of an NPC as well as the mood of the game space, much as Smith contends in his discussion on emotion cueing,[6] the facial expression itself must provide the emotional clarity the game designer intends. If the goals of characters steer their action, then the emotional expression on the face of NPC's reveal how the designer wants the player to understand the character's attitude toward their goals and the consequence of their action at a particular instance. (Later in this discussion, there will be a more robust elaboration of how Computational Psychology informs the design of the internal workings of NPCs.)

But is there enough information from the face of an NPC to "move" the narrative and cause the player to decide something significant within a game? The answer may in part depend on the resolution of the animation. A higher resolution facial animation provides the character designer with more expressive possibilities. In the past, the concept of visual resolution has described the density of pixels for a given square of screen surface. But *animation resolution of the face* describes the number of vertex groups of an NPC's head mesh that move distinctly and simultaneously for a unique emotion expression. While this description does not yet provide a quantifiable ratio of dependent and independent variables (like polygon counts, for example), it

does communicate a problem in game animation that affects the player's experience. Game animation must render each frame in real time (usually 60 frames each second). A character with complicated expressions simultaneously moving many parts of a mesh, can dominate the parallel calculations of a graphics rendering process, thus challenging the graphics processor to complete all necessary calculations within $1/60^{th}$ of a second. Cinematic acting for NPCs depends on such computational power because many vertex groups that simulate facial emotion expressions must move in consort. But with complex animated facial expressions, players can observe more detailed expressions; they can read more nuanced emotional information when observing NPCs up close. Chatman's close reading examines the polysemic phrasings of Maupassant's story with the polysemic cinematic expression of Renoir's filmed rendition. High resolution animation allows a similar comparison between a character in a work of literature and its representation in a video game adaptation. Complex facial expression in game characters can allow the player to perceive NPC emotions in ways similar to human actors. NPC facial expression can then be used to effect narrative generation.

Augmenting the expressive capabilities of NPCs has the side effect of augmenting narrative expressivity of video games. Narratologist Uri Margolin defines *characterization* as a "human or human-like individual… capable of fulfilling the argument position in the propositional form DO(X) – that is, a Narrative Agent (=NA), to whom inner states, mental properties (traits, features) or complexes of such properties (personality models) can be ascribed on the basis of textual data"[7]. The definition is applicable especially when examining the detailed form of a game character as seen in a single frame of video game play, in a model sheet typically produced in the preproduction stages of character design, or in a written biography of a video game character. In these video game preproduction documents, signification occurs by inference.

The character's behaviors are imagined by what its body looks like statically, and what has been reported about its past in the written form. Margolin continues by describing *character-building* in a narrative, which is an "accumulation of a number of traits" or trait clusters "from several successive acts of the NA [Narrative Agent]…" A built character is "a unified stable constellation" of traits experienced in narrative time. While Margolin had in mind primarily literary characters, the player likewise experiences NPCs during gameplay over time and observes their actions in much the same way, gradually building an idea of a character in mind.

When one thinks of actions in video game play, often the idea conjures dramatic movements that propel the body of the character or impact objects the character touches. These actions reveal internal and external states and features of NPCs through body gestures, postures and facial expressions. Facial expressions tell much about the internal mental states and actions of NPCs. Literary representations of emotion take the form either of direct predicate statements where the emotional state becomes the predicate or adverb (e.g., It is angry, charmed or, It acts lovingly, scornfully) or are presented indirectly as actions or reactions in the form of a verb (e.g., It *flees* from fear, It *flirts* from arousal). Visual representations of character must use other means. Posture, gesture and facial expressions of emotion are the primary means of presenting emotion in NPCs. Early video games, much like early animation in the cinema, were limited to a smaller vocabulary of emotional expressions because the processor speeds and workflow did not easily allow for much variation of expression. But more powerful graphics processors and modularized workflow allow for complex groups of simultaneous animations of different parts of an NPC face to form many more combinations and variations of emotional expression. By expanding the vocabulary of a character's emotive expression, the possible meanings derived by the player from an NPC's expressive system also expand. With more possible meanings entering

the space of a narrative, the count of possible player experiences and interpretations grow as well. Thus, as NPC emotional expressivity increases in combinatorial complexity with an expanding vocabulary of movements and gestures, NPCs can evolve from expressing simpler cartoon-like acting styles where performative meanings are simple, discrete and the expressive variables show few dimensions, into a more robust cinematic acting style where expression is polysemic, often ambiguous and the expressive variables of the face have many more dimensions.

2.3 Agency and Acting

The "construction" or "building" of NPCs by the player within a video game borrows heavily from experiences with other non-ludic art forms, such as literary fiction and films, as well as perceived experiences of simulation systems, such as software. Theatrical and cinematic narrative depend on the appearance of willful actions of characters in a story world. *Agency* in drama is the apparent freedom of a character to act, react, not act, or sublimate desire to achieve a goal within a fictional world. Aristotle and many modern theorists of agency attribute voluntary goal acquisition as an underlying cause of human action[8]. For NPCs and Player Characters, Janet Murray translates the concept effectively. "Agency is the satisfying power to take meaningful action and see the results of our decisions and choices".[9] The definition of agency has extended itself into software engineering.

As NPCs in video games developed, likewise *agents* were run-time entities within software. The concept of *agency* in software engineering emerged commercially in the mid-1990s when software required run-time decisions to manage its states autonomously. Developers realized that software systems need agents to assess the state and behavior of a system's environment to autonomously manage the system in the background so that desired measurable

states are maintained. Agents in software are distinct from other software components because they do not need to wait for a user to tell them what to do. Agents can have goals that they remember pervasively, and they can be provided readable and writable access to execute methods on other software objects to accomplish those goals.[10]

Eventually, agents within software surfaced to interact with users directly using text, speech or character animation. They helped users accomplish software dependent tasks. These have been called *virtual* agents. The primary functional tradition of servitude for virtual agents limits their autonomy and their range of emotional expressivity. Virtual agents were not designed to entertain. But as computer games began to use facial animation for emotion elicitation of NPCs, an emotion model architecture was required. Animators provided basic iconic facial expressions of emotions and software developers integrated morph animations driven by emotion models adopted from virtual agents.

Virtual agents, however, were not originally embedded in artfully crafted interactive narratives, but were relegated to software assistance for users. As emotion models of Cognitive Psychology were implemented into virtual agents, validation of virtual agent efficacy affirmed their use for pedagogical applications.[11,12] The implementation of the same virtual agent technology for video games was an obvious temptation. But the limitations of graphics processors, central processors and memory demanded a low bar of aesthetic expectations for players as demonstrated in the crude character designs of the first three decades of computer games. Player aesthetic expectations of autonomously animated NPC “acting” had to be simpler to adapt to these technology limitations. Following comic book artists and non-photoreal animators before them, game character artists and developers more often mimicked the expressions in the cannon of Delsartean-based typology of facial and body models to create

automated animations that could express basic recognizable emotions.[13] Virtual agent emotion models could at best validate their "ground-truth" assumptions based on a player's recognition of an NPC emotion, an NPCs believability of the authenticity of expression, or the NPCs social appropriateness. These three assumptions reflect the narrow range of expressivity a virtual agent could emote given the technological constraints on real-time 3d animation. These methods of emotion expression in agents do not consider the concept of complexity of animation expression, which can be defined as the number of simultaneous facial animation vertex groups simultaneously activated that an NPC may elicit in response to stimuli as either an active or reactive expression. Simultaneous animations with higher resolution of simultaneous movement allow developers and animators to implement more complex emotion elicitation in NPC behaviors.

2.4 Behave Like a Human, Think Like a Machine

MIT Media Lab researcher Patti Maes in 1997 wanted agents of "intelligent software" to "sense the current state of its environment and act independently to make progress toward its goal".[14] While describing agents, Maes uses one of the underlying concepts of artificial intelligence laid out forty years before by Alan Turing in his notable 1950 essay "Computing Machinery and Intelligence".[15] Turing proposed that machines could think autonomously and learn from their environment and past actions. But four decades earlier, Turing went further than Maes. He proposed that machines could respond in human-like ways to elicit enough rapport to convince a user that the machine was human. Without stating so directly, his rebuttals redrafted the idea of what "thought" was by allowing that the production of a thought can include distinctly "mechanical" methods. The notable proposal of a continuous state machine suggests that he at least intuited some types of "thought" were not finite states that depended on a combination of

true (1) or false (0) relations but were gradients that allowed for a continuous range [0,1], much as basic emotional expression states are currently given normalized values, and much as emotion expression morphs are blended between each other (e.g. blend shapes in 3d programmed animation in video games). Turing anticipated that a machine may have to perform machine-like thinking to come to human-like results. “May not machines carry out something which ought to be described as thinking but which is very different from what a man does”?[16] Virtual agents, as Turing described in his essay, may process “thought” like a machine, but still elicit behavior like a human. Virtual agents are programmed to elicit emotional reactions like humans while processing input data mathematically, probabilistically and algorithmically.

Long before computers existed, French philosopher and mathematician, René Descartes reasoned how a machine could never think like a human.[17] Descartes decried in 1668 the “impossibility” of sentient AI. “...It is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act”.[18] The reasons he gave described the challenge of computing semantic correctness and social appropriateness with limited computing power and memory to process a wide range of meaningful expression combinations.

What Descartes required as basic needs for a machine, semantic correctness and social appropriateness, are required skills that humans also do not always do well: *teaching* and *entertaining*. Both human activities require sensitive and dynamic awareness of the emotional state of the user. A detailed psychological model of the user or player is essential to direct the actions of an agent precisely and to measure the agent’s effectiveness. A virtual agent with eyes must know where on the body of the elicitor it should look and gather input to make decisions or to react. It should also be able to interpret what that input means. Gestural and facial expression,

with its potential for producing ambiguous meanings that trigger emotions, is what Descartes intuited as a combinatorial morass.

2.5 Pedagogical Agents and Emotional Behavior

Descartes was correct. Logical processes are not effective to make decisions if available information is not complete enough to find a rational conclusion, or if premises are contradictory. Cognitive psychologists and neuroscientists contend that "emotional intelligence" allows the mind to fall back on another process that is possibly based on probabilistic instead of deductive reasoning, and what may be the function of "basic emotions".[19,20] The brain contains neural networks that use pattern recognition to provide probabilistic answers to situations that rational thought cannot adequately answer. By applying similar emotional reasoning methods in a virtual agent, it can "predict" or recognize patterns from machine-learned (ML) instances to make behavioral decisions.

2.5.1 Socially Appropriate Behavior

Rosalind Picard, a research pioneer in the field of Affective Computing, describes an emotionally intelligent agent as possessing "abilities to recognize, express, and have emotions, coupled with the ability to regulate these emotions, harness them for constructive purposes, and skillfully handles the emotions of others".[21] In the context of an animated video interaction, a three-dimensionally animated agent often moves and speaks in ways perceived as "socially appropriate" for the context of the user-agent relationship. It should demonstrate emotionally logical behavior, follow an expected cause-and-effect sequence, that would respond to the sentience of the user. The term socially appropriate (which is sometimes called "behaviorally appropriate") describes behaviors that fall within a normal range of variance of behaviors given a specific social situation with a known set of persons in specific types of relationships to a user.[22]

The range of variance is affected by an expectation of satisfaction from a user and the culturally determined role the agent plays in the relationship with the user. For example, a virtual agent classmate in a learning application who assumes the role of a *peer* will likely behave with a broader range of appropriate behaviors than a virtual agent that is an *instructor* who assumes a role of pedagogic authority. Behavioral variance in virtual agents can constrain or expand user input as users adapt their decision-making variance to the variance expressed by agents.[23] Additionally, some attribute values of the human user are considered that can affect an agent's behavior – physical location, language, culture, gender, class, education level, age, sexual orientation, country of origin and ethnicity.[24] A virtual agent can be designed to react to vocalization variables: diction, prosody, volume, pitch and backchannel utterances or for facial expression variables: movement direction, velocity and acceleration so that it can reflect and respond in socially appropriate ways to the user. A designer's intention is for a virtual agent's elicitation parameters to effectively tune to the user's familiar behaviors so that it will maximize rapport during dyadic conversational interaction.[25]

Virtual agents that teach or entertain challenge the engineer to model their algorithms to accomplish a task that has no finite outcome of "correctness". A virtual agent with affective abilities should at least be able to fulfill each of Picard's criteria for emotional intelligence – emotion recognition, elicitation, and regulation – to accomplish its goal. For example, a virtual piano teacher must demonstrate a nomenclature in musical theory, but also must sense (by facial or vocal expression) if the student "feels" she is able to understand more complex or simple examples, or no examples at all.[26] The virtual agent must decide if its own gestural and aural expression should be firmer and more resolute in posture and tone, or gentler and more accepting. In one example, a hospital-based virtual agent, called a *Hospital Buddy*, provides

companionship and reduces psychiatric side-effects of long-term patients. The *Hospital Buddy* must sense a level of trust to inquire from the patient perceptions of his hospital experience.[27] Virtual immigration interviewers in the E.U. and the U.S. have been developed and tested for border crossings.[28] They are designed to flag possible deception by visa applicants using affective biometric data, including facial micro-expressions, during inquiries for facts collected from an interview. The agent must choose questions that intend to trigger emotional responses in relation to known facts and to notice an absence of emotional response.[29] In all these examples of pedagogic virtual agents, they must sense the emotional state of the user and express itself in a socially appropriate way so that its tasks and manner prioritize the productive flow of information between computer to human, and back from human to computer.

2.5.2 Elicitation by Design

For this discussion here forward, the term *pedagogic* agent shall refer to didactic, interrogative, informative and simulative virtual agents (all characteristics of a teacher). Pedagogic agents are generally designed to elicit "controlled" emotional responses so that the messages they express produce a predictable experience from the user and for the developer. For pedagogic agents, emotional expression must carry an intended message and outcome so that the user's behavior leads to an interactive experience that enables the agent to gather the emotional information from the user as intended – surprises can defeat the software's purpose. Pedagogic agents are not free to express the full capacity of a human personality any more than a human could doing a similar pedagogic task. To accomplish this emotion-elicitation-by-design, the author of the agent must program to specification by biasing the emotion that the agent should demonstrate during an instance of human interaction. Pedagogic agents for the most part substitute a human service that the user is presumed to need. Once the user ceases to need the service, then a pedagogic agent

can become an intolerable sycophantic nuisance. The "coached" emotion-elicitation-by-design approach often fails to convey authenticity because the emotional "results" performed in a scripted interaction are intended to synchronize with a developer-permitted pedagogic task to fulfill an anticipated user need rather than respond to the spontaneous emotion elicitation of the user.

2.5.3 "Negative" Personality Traits

Unpredictability, Immorality, Deceptiveness, Self-Loathing, Flippancy, Distractedness and Forgetfulness. These seven words are negative personality traits discouraged for any person assigned to a pedagogic role. Teachers, counselors, judges, doctors would not likely gain more clients if user reviews included descriptions with these negative characteristics. However, they are all attractive characteristics for fictional NPCs because they can create inter-personal behaviors where contradiction and conflict reveal and belie the moral problems of humans. Encountering troubled characters tests a player's beliefs and assumptions about the game world (and perhaps the player's world) in ways that make game play pleasurable. The player often seeks to resolve these contradictions and conflicts in their game world as they are embodied and sometimes resolved in characters. (Think of the puzzling behavior of Darth Vader in the Star Wars franchise as he seems unable to deal a death blow to his adversary, Luke Skywalker. The viewer endures a decade of patience and multiple sequels to finally hear Darth Vader's revelation: "I am your father."[30]). While these characteristics may not be desired in a non-NPC emotion model (such as that of a virtual agent), each of them makes useful functional components in an NPC emotion model as it decides either what emotion its appraisal will select, or what elicitation it should animate.

2.5.4 The Risk and Pleasure of Neuroticism

Sometimes the player appreciates hints of awareness of the negative characteristics in human behavior, or what is typically referred to in personality dimensions as "neurotic behavior".[31] Neurotic behavior from pedagogic agents (virtual or real) distract from the disingenuousness of a commercial-cum-social interaction. (Consider the cynical humor of flight attendants who joke about passenger attention during pre-flight announcements.) But neuroticism tolerance varies among users and contexts; it may contradict the social "believability" of a virtual agent.[32] While humor for the user of agent-enriched application can be the result of an agent's neuroticism, it can be socially inappropriate for a user-agent relationship if the agent is pedagogic. Additionally, if the pedagogic agent has little or no awareness of human misdeeds or maladaptation, then a neurotic response would be implausible. Moreover, a pedagogic agent must have a narrow range of expressive variability to keep the user serviced and focused on the tasks of the software service.

2.6 Distinct Emotional Features of NPCs

An entertaining agent gathers, processes, and elicits sentient information with greater freedom of emotional expression than a pedagogic one, much as an actor has more emotionally expressive liberty than a teacher. *Entertaining* virtual agents are found in games and interactive media where the agent is personified to express the features of its personality that serve the game world. For clarity, we refer to entertaining virtual agents as NPCs. Their emotion-response algorithms receive relevant data from the player and game play space and elicit expressive data back into the game play environment for the player to observe and respond. A cycle of data exchange can create rapport or dissonance between player and NPC as the game evolves and as the NPC's role requires. The narrative "facts" revealed in this exchange can be distorted or omitted by the

NPC's subjective filters, much as any fictional character does in other media. Furthermore, neuroticism may be welcome insofar as it fits the role required for the world of the game. NPCs are often designed to elicit from the player a wider range of emotional responses than pedagogical agents. Similarly, live action actors and comedians or animated cartoon characters provoke emotions that are not permitted from instructors or public spokespersons. An actor, through his character, can elicit rage, fear and disgust in "socially inappropriate" ways when playing a conflicted character. Pedagogic agents such as piano teachers or public health officials usually cannot; the roles of pedagogic agents rarely allow neurotic behavior. Spectacular neuroticism as a performative aesthetic of cinematic acting is a feature of cinematic acting insofar as game players also watch movies and expect a similar degree of performative complexity, emotional authenticity and celebrity attraction.[33, 34]

2.6.1 Cinematic Acting Style

The trend in the last decade has demonstrated strong interest in cinematic acting styles as games increase their ability to use higher resolution graphics, motion, and complex facial expression animation.[35,36] Cinematic acting differs from theatrical acting in the production method. While performance preparation methods are very similar, cinematic acting requires the actor to deconstruct a performance into smaller units for each shot. Acting for animation of video game NPCs requires actors to decompose further – down to the gesture, posture, or facial expression. To capture motion data for high resolution animation, increased computational power is needed to recognize the face, track its inflection, and predict its movement over time. Contemporary graphics processors allow for the capture of more complex and simultaneous emotion expressions. High-resolution animations support the aesthetic of cinematic acting styles in video games.

What are cinematic acting styles? Cinematic acting inherited many methods for training from theater and screen media. There are two dominant currents in cinematic acting styles. One works from a collective external vocabulary of postures, gestures and facial expressions that are loosely assigned to roles within a dramatic or comic narrative. The task of the actor is to shape their own variation of a role as a set of expressions that convey the emotional expressions of a character for the specific narrative. This approach, often called the "outside-in" approach and is associated with performance theorist Vsevelod Meyerhold and Francois Delsarte, is highly adaptable between cultures and requires a studious actor to learn the emotive vocabulary of the target audience. The outside-in approach affiliates well with animation and its tradition of referencing a Delsartean taxonomy of physical expression, much like trained and practiced for silent film acting.[37]

Another major approach requires a process of self-examination to identify existing emotional impulses from past lived experience so that when an actor in-character focuses on an action, the resulting emotion from obstructions or assistance are personally authentic. This inside-out approach has been affiliated with Stanislavski and his Method approach as well as his Meisner. In animated or live-action films, both currents of cinematic acting deploy complex simultaneous movements motivated by inferred psychological states that are used by actors to create character facial expression for motion capture or voice recording for video games.

For this research, both cinematic acting styles provide valuable methods for facial expression in video games as they connect the emotional source of actor expression to NPCs that move narratives forward. Games that aspire to cinematic high resolutions of image and movement, such as *Beyond Two Souls*[38] and *The Last of Us*[39], have shown substantial consideration of the aesthetics of cinematic acting through complex animations of the face.

Therefore, the methods and techniques of cinematic acting, character development and performance deserve consideration as video games embrace NPCs with cinematic acting styles.

2.6.2 Responsiveness to Multiple Simultaneous Stimuli

The game player, attracted to cinematic acting, expects more complex animation to enrich the perceived emotional and intellectual life of the character. To develop cinematic acting for NPCs, its animation must move in complex ways that are consistent with the design of a complex screen performance; this means moving groups of vertices of an NPCs mesh in ways that represent a response from simultaneous stimuli from both external-physical and internal-mental sources. When we closely examine well-prepared actors that move in any fully developed performance style, such as the naturalism that the Method acting training espouses or physical comedy, one property that makes a performance complex (and possibly "good") is when an actor responds to multiple stimuli almost simultaneously, from sources that are external-physical, and from sources that are internal-mental. Acting theorist Lutterbie draws from Cognitive Science as he describes the "acting instrument" as a Dynamic System Theory model that listens inside and outside the body, filters decisions through memory, then responds with complex movement, language and gesture.[40] Complexity increases when an actor listens, acknowledges and responds with "executive control" to multiple stimuli with an intensity that seems appropriate to the scene, role, its story, and its world.

Often attention to stimuli in the external-physical space appears to the player with a subjective internal-mental amplification or diminishment of intensity. As game designers deploy cinematic acting styles, the Player observes stimuli generated from the NPC that seems internal-mental and occurs simultaneously with stimuli from the external-physical space. That disparity of intensity creates a competition of focus for the NPC which results in polysemic expression.

The player notices this distinction and imagines, often with pleasurable uncertainty and insatiable curiosity, what the actor is thinking or feeling. Often the most pleasure-inducing experience is an ambiguous facial expression, and the audience or player invents or projects their own meanings that caused the expression. This ambiguity of focus can be built into the NPC and is fundamental to how cinematic acting styles contribute to the expansion of the narrative space. No longer is the game limited to a visual-physical space. Cinematic acting styles open the narrative to the mental space of the NPC. To achieve cinematic acting, NPCs require uniquely subjective filters and algorithms to idiosyncratically appraise and elicit reactions to the emotion-evoking entities it perceives.

2.6.3 Cinematic Mediation of Expression

Principal game characters are often presented to the prospective player base with spectacular trailers and posters, much modeled on theatrical film trailers. These advertising media reinforce character archetypes and stereotypes, while also creating a set of behavioral expectations in the player before the player ever plays the game. Additionally, in-game NPCs are experienced by players as mediated persona through game video screens and speakers, much like film and television characters. Audio-visual moods and cues reinforce mediated emotional expressions in a game much as one observes in film and television.[41,42] The construction of mediated emotional expressions results from already familiar production methods of actor training, casting, rehearsal, recording and reperformance technologies that games use for their production. A facial expression on an NPC after a kiss to the lips or bullet to the back triggers an expected category of emotional response that can stand alone or come accompanied with sympathetic or dissonant music, sound effects or dialog. The combined facial expression and sound can affect the perceived meaning of the response and negate or affirm the expected response. These

expectations do not precisely resemble real life but are a synthesis of memories of watching previous mediated experiences of similar events combined with real life memories. Actors are trained to express real emotions and to synthesize them with culturally preconceived expressions of real emotions.

2.6.4 Performative Mediation of Expression

Mediation is not only determined by industrial or societal norms through cinema. Mediation can also be the result of an actor's own imaginary construction of the character's mental processes as they believe the audience will perceive them. As each elicitation of an actor or NPC is presumably the result of the flow of emotion-evoking stimuli, it is an actor's detailed design of those mental processes for the character that provides actor-centric mediation of elicitation. Thus, what has here been defined as cinematic acting style for NPCs, with its complex mesh animations, is the result of an actor's invention of an internal emotion model that filters sensations, identifies them, prioritizes them, associates them with memories, maps sensations to affect definitions, and selects an elicitation in response that appears as an emotional facial expression.

2.7 Computational Emotion Modeling of Cognitive Psychology

Performance theories that are foundational for cinematic acting share some concepts that describe the derivation of human emotion with those defined in the literature of the computational field of Cognitive Psychology. Psychologists Ortony, Clore and Collins were among the first to propose a human emotion model for simulation on the computer. While theirs was not intended for game development, it is easy to imagine that its deployment would sense emotion triggers in the game play space. The model would process the sensed data through its own character-specific filters, choose an emotion that describes the state of the NPC, and then

elicit that emotion through an animatable face, body or synthetic voice. During the last thirty years, emotion models were designed to test the human affect processing system by observing them embedded in virtual agents more than NPCs. Cognitive Theories of emotion are collectively known as Appraisal Theories. To understand the rationale for an Actor-Centric emotion model for NPCs, a fundamental understanding of an Appraisal Theory framework for emotion models is helpful.

2.7.1 A Brief Summary of Appraisal Theory for NPC Emotion Modeling

Cognitive Psychologists for the most part agree that emotions are generated in response to an "appraisal" of the environment and its condition to fulfill and obtain one's goals (however grand or modest), affirm one's behavioral *standards*, and satisfy one's *tastes*. A person's environment is often laden with entities to appraise in ways that concern a person's physical and social survival and quality of existence. Appraisal occurs initially as a person assesses the immediate situation and its immediate impact.[43] And for some theorists, there is a *secondary* appraisal (reappraisal) where one considers how to cope long-term by accepting the situation as it is and adjust one's attitude about the situation. A person might then take action to change the situation in response.[44]

Entities that have emotional effects on a person are either inanimate *objects*, animate *agents*, or consequential *events*. A person will judge an object's *likeability*, an event's *desirability* (of consequences), or an agent's (past, current or future) *praiseworthiness* (of actions) in terms that either favor positively or negatively one's own goals, standards and tastes. Observing and evaluating an entity (object, event or agent) and its feature values registers an emotional appraisal along one or more dimensions of measurement. The values of these measurements are stored as *appraisal variables*. A fundamental scale of measurement is

pleasure and *displeasure* within a range of [-1,1]. This dimension has also been interpreted as *aversion* and *attraction.* As a two-dimensional appraisal scale, *pleasure-displeasure* runs on the X-axis and rises with *arousal (intensity)* on the Y-axis within a range of [-1,1]. Originally known as the Circumplex Model of Emotion,[45] it has been further refined as a 12-emotion unit circle with evenly divided expression positions. This variation has been named by some psychologists and neuroscientists, Core Affect,[46] as shown in Figure 2.1. Core affect prioritizes the "basic" emotion positions in the unit circle. A three-dimensional scale variation includes *dominance* on the Z-axis within a range of [-1, 1],[47] where dominance represents the degree to which a person can control the emotion from triggering impulsive elicitations that could be anti-social, unwelcomed or inappropriate, as shown in Figure 2.2. Most Computational Cognitive Models based on Appraisal Theories use either two- or three-dimensional emotion mapping, as found in Embodied Conversational Agent development.[48] Using two- and three-dimensional emotion maps is known to provide a system modeled on closely observed human behavior. Emotion maps allow researchers to test a virtual agent accurately in relation to a human reference.[49, 50]

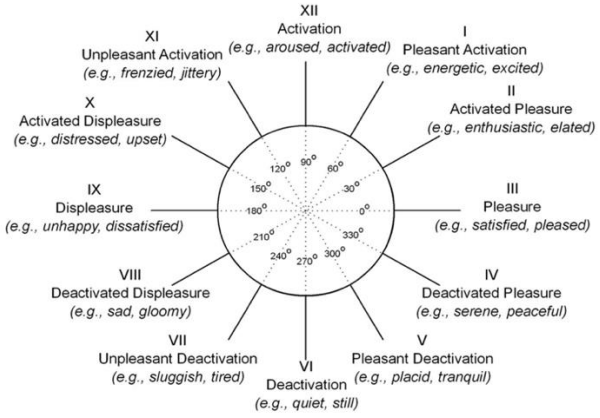

Figure 2.1: Core Affect Circumplex. A two-dimensional emotion map that divides a unit circle into 12 sections of "basic" emotions.[51]

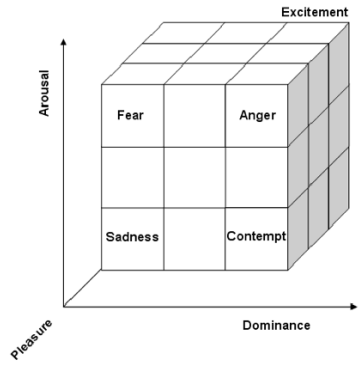

Figure 2.2: Three-Dimensional Emotion Map. PAD to organizes emotional categories for a study on glances.[52]

NPCs perceive entities with subjectively determined values of their feature variables based on their character-specific goals, standards and tastes. For every entity that triggers an emotion, an algorithm must translate and map all the entity's feature values in a two-dimensional Circumplex, as shown in Figure 2.3, or a three-dimensional volume, as shown in Figure 2.4, to a coordinate point positioned by its appraisal variable values -- Pleasure, Arousal and (if in three-dimensions) Dominance (PAD). These two- or three-dimensional values are normalized within a [-1, 1] range of a unit circle or sphere.

Much of the distinction between appraisal theories is the structure of the emotion model, and therefore the path of stimuli signals through the model. Some theories emphasize memory and pattern recognition components in relation to goals and praiseworthiness, others emphasize reappraisal and changes of emotion response, while others emphasize the effect of stimuli on planning for goal acquisition. The resulting PA or PAD values may change as the appraisal process cycles through each appraisal or reappraisal over time. PA and PAD values infer a vector with a coordinate point at its tip. With each reappraisal cycle, the vector can move to a new position as the emotional state of the agent changes.

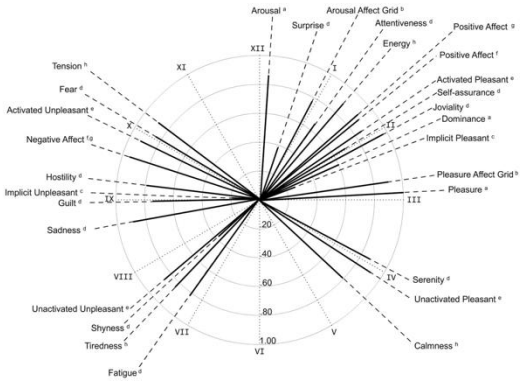

Figure 2.3: A 12-Point Affect Circumplex (12-PAC). Model of Core Affect to plot the vectors of 30 mood words.[53]

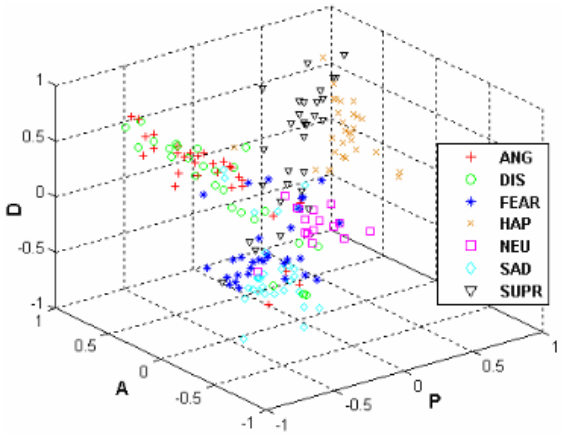

Figure 2.4: Three-Dimensional Emotion Map. Used to plot facial expressions of a talking avatar from an expression database.[54]

A coordinate point within a unit circle or sphere falls into a region or volume that is bounded by an emotion label. The idea of the emotion label is controversial. Some Cognitive Psychologists prefer to abandon labels for emotions. The trepidation to work with emotion labels stems from the subjective quality of assigning a culturally determined word to an object (an emotion in the body) known by its elicited effects, and often repressed by the vagaries of culture.[55] Labeling an emotion is a linguistic problem rife with semantic contradictions of inclusion and biases. However, since NPC design depends on collaboration among creative and technical professionals, labels must be used to communicate software development processes.

At the top of a hierarchy of emotion labels are "primary" or "basic" emotions. Some psychologists believe that these are triggered by biological instinct and evolutionary necessity.[56] If the label is not a primary emotion, then it is a secondary emotion. Another branch of the psychological and neuroscientific disciplines disputes this concept of "primary" or "basic" emotions. Those in disagreement contend that emotions are neurologically constructed after birth as the brain forms its emotion generation system in response to the social and physical environment. These Cognitive Psychologist and Constructivist Neuroscientists believe that many emotions are contextually constructed, categorized and classified as expression concepts, and that not all expression concepts are equal from one culture to another. Constructivists contend humans have a proclivity to express many emotions only if learned and permitted, and that some expressive characteristics may not be innate at birth.[57,58,59] The labeling of patterns "nurtured" in context, explains the differences in emotion label meanings from one culture or social group to another. A more detailed elaboration of the distinctions in emotion theory as they related to modeling NPCs and annotation of video corpora is found in a later chapter of this research.

2.8 Solutions and Recommendations

For an NPC emotion model to accommodate the complexity of perceiving and processing multiple emotional entity categories of stimuli simultaneously, components that sense and process emotion-evoking entities must be built into the emotion model for sensing, evaluating, prioritizing and selecting affect. Ultimately, the path of either a sensation signal or a processed affect selection is controlled by algorithms that schedule the flow of data from sensation to appraisal to elicitation so that behavior is synchronized to receive new sensations and to reconsider past sensations (reappraisal) held in memory. Present tense sensations and memories compete for attention. Simultaneous appraisal and elicitation with reappraisal benefit from parallel processing and scheduling to manage the complexity of these computational processes. Cinematic action and therefore high-resolution autonomously animated emotion expression use such complexity in an emotion model and must increase with the growth of more specific components and more non-linear scheduling of affect selection and elicitation.

2.8.1 An Actor-Centric Emotion Model

Extracting the principal points identified thus far, an actor-centric emotion model should have components that enable an NPC to optimally elicit movement of the face. Arrays, structs or vectors may serve to store mutable and immutable ordered data that describe the implicit biases of an agent. But a Feed-Forward Neural Network (FFNN) will provide the structure to algorithmically process input from multiple handlers-listeners or sensors into a gradient output that will result in an animated elicitation. Table 2.1 associates a Performative Characteristic with an appropriate Data Structure and corresponding Algorithmic Steps. These associations are found in a selection of computational emotion models of the past that we will discuss.

TABLE 2.1: DATA STRUCTURES AND ALGORITHMS FOR PERFORMATIVE CHARACTERISTICS

Performative Characteristic	Data Structure(s)	Algorithmic Step(s)
Expression of Negative Traits	Arrays of vectors that store coefficients to alter values of post-appraisal dimensional vector	Receives post-appraisal signal as dimensional vector. Each dimensional value is multiplied by a coefficient $C \in \mathbb{R}$
Multiple Simultaneous Sensations	Multiple FFNNs or structs with object handlers-listeners and feature recognition array	FFNNs or structs receive signals from environment during same frame (appraisal instance).
Internal Objects	*Objects in NPC:* A first array stores feature values of significant objects from previous appraisals (memories). A second array stores relevant structures indexed by object types whose methods can trigger internal sensations.	If external stimuli are recognized as similar to significant objects in first array, then method in structure of matching object type in second array triggers internal sensation.
External Objects	*Objects in game space:* Struct with vector representing feature values. Objects in player space: FFNN recognizes object and converts to vector representing feature values.	External stimuli are recognized as similar to objects in first array, then method in structure of matching object type in second array triggers internal sensation.
Self-Oriented Goals	Data types that permit gradient values, such as multi-D vectors. Any variable if the goal is not an emotional state or change in emotional state.	Method must check goal value and current state value. Difference may trigger new appraisal. Method might check player variables to decide to forgo action to serve self and instead engage player.
Modularity	Affect Derivation Component object or struct contains arrays or links to FFNNs. More modules increase ability to recognize more emotional entities.	Method switch can enable or disable modules as needed to minimize computational expense as agent changes state or space.
Scalability	Within Affect Derivation component, increasing affect components allows for more granular sensitivity to variations of entity feature variables.	Method switch can enable or disable components as needed to minimize computational expense as agent requires greater sensitivity to entity feature variables.
Affective Loop	The entire emotion model runs as a two-cycle process. *Cycle 1:*	Scheduling of signal vector must allow for each component to

	entirely internal and allows reappraisals. *Cycle 2:* partially external to the agent, allows the agent to send out elicitation and receive response from other agents or the player.	complete its task. Must store in memory significant appraisals for subsequent recognition or reappraisal. Must regulate and prioritize competing signals to determine which should occupy appraisal bus: new signal or reappraisal.

A generic emotion model as shown below in Figure 2.5, presents basic components of an NPC. This generic model was synthesized from the MAMID model developed by Hudlicka,[60] from the FeelMe model developed by Broekens and DeGroot,[61] and from the FAtiMA model developed by Dias, Mascarenhas and Paiva.[62] The resulting model satisfies Picard's componential requirements. Table 2.2 elaborates on the function of each component.

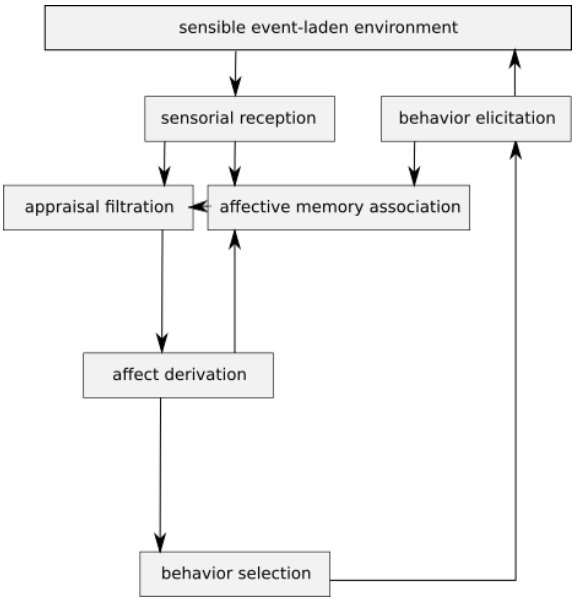

Figure 2.5: A Proposed Emotion Model. Synthesizes features from Hudlicka's MAMID,[63] Broekens and DeGroot's FeelMe,[64] and Dias, Mascarenhas and Paiva's FAtiMA models.[65]

TABLE 2.2: COMPONENT DESCRIPTIONS OF THE GENERIC EMOTION MODEL IN FIGURE 2.5

Component Name	Description
Sensible Event-laden Environment	Container of all events, objects and agents that can be sensed by emotional agent.
Sensorial Reception (sensors)	Events that are accepted by sensors.
Appraisal Filtration (filters)	Filters perception of objects, agents, events in terms of goals, standards and tastes.
Affective Memory Association (data storage)	Storage of perception, affect, and elicitation of self and others. Creates memory networks and evolves rules from associations so as to optimize rewards toward fulfilling goals, meeting standards and satisfying tastes
Affect Derivation (emotion values assigned)	Pleasure-Arousal (-Dominance) mapped emotion vector algorithm.
Behavior Selection (select blendshape)	Chooses an animation of the face indicating current emotional state.
Behavioral Actuator (animate blendshape)	The physical component that generates movement of the mesh as gesture or facial expression.

Inside of the Affect Derivation component in Figure 2.5 is what is depicted in Figure 2.6. The generic emotion model uses aspects of FAtiMA's modularity.[66] The Appraisal Components provide the Appraisal Frame the current values of each Appraisal Component's predictive result. We propose that each component consists of an FFNN that recognizes the entity and its emotion relevant feature variables and values to predict appraisal variable value for each affect from [0, 1]. For a facial emotion expression, that result would be the gradient value of an emotion expression of which there is usually at least six expressions plus neutral.

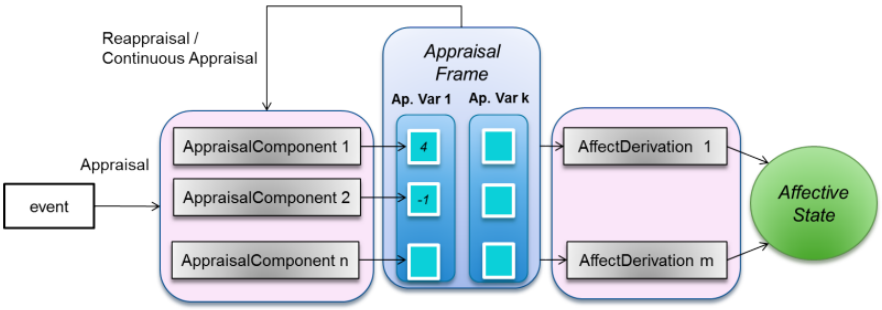

Figure 2.6: Detail of Affect Derivation Component. From FAtiMA emotion model.[67]

Each Affect Derivation Component consists of a data structure that receives input from an Appraisal Filter. The Appraisal Filter receives an appraisal signal that recognizes each emotion expression for each respective Appraisal Component. The filter will likely implement an object recognition neural network or an array of feature value ranges – both will process a vector whose elements and their values are within range to process and predict a meaning. For a facial emotion expression system, the Appraisal Filter will look for facial features in a video stream that fit the parameter values of an emotional expression on a human face. Then the signal must be sent to the appropriate Appraisal Component within the Affect Derivation unit. An Appraisal Component will evaluate a feature of the face that best indicates a value of a specific emotion expression type (happiness, anger, etc.). Then the Appraisal Component must use an emotion recognition FFNN to predict the expressive meaning of the signal. The resulting gradient value will reside in the Appraisal Frame. All Appraisal Frames will be evaluated to choose a final Affect Derivation for the appraisal instance. The resulting derivation will be the Affective State, which will make a Behavior Selection, and in turn trigger a Behavior Elicitation.

2.8.2 Applying the Affective Loop for Emotion Model Dynamics

A consistent theme in the discussion of emotion models is the idea that affectively relevant sense data runs a cycle between an agent and its environment. This characteristic is found in the last row of Table 2.1. Sundström calls this cycle the Affective Loop,[68] and it is complex enough to deserve a more elaborated explanation. The Affective Loop has three stages familiar to actors from improvisation and rehearsal exercises: (1) a human user elicits an emotional expression, (2) a system responds with a responsive emotional expression, (3) the human user internalizes and interprets the affective meaning of the eliciting system, and (4) the loop repeats. A similar concept is found in the Repetition Exercise of renowned theater director, acting teacher and theorist Meisner.[69] The three stages of the Affective Loop have parallel stages in the process of developing an actor's skill at finding authentic emotional responses to interactions with other characters, objects and events observed in a scene. The Affective Loop expands human-to-human interaction, as practiced in the dramatic arts, to a "design principle" of the Affective Computing domain that implements motion-capture-performer-to-computer and computer-to-player principles.

The Affective Loop design principles are interpreted for player-to-NPC interaction as follows from Sundström. *Embodiment* is enhanced as the player's body expresses its affect through the game controller or other sensing interface. Some part of the player's elicitation data becomes input that the NPC's emotion system can sense. Likewise, the player can sense the NPC sensing the player, and sensing the NPC eliciting its organismic idiosyncrasies through its own body.

Flow is another attribute where, as the player interacts, there is a sense that there is always potentially a new sensation coming from which to repeat the Affective Loop. Or, an NPC

can react to internal stimuli from memory. The emotion model not only repeats as loops, but it is normally open for new stimuli from internal or external sources that allow each loop iteration to elicit distinct values.

Another Affective Loop attribute occurs when NPCs are designed with elicitation *ambiguity*. When there is more than one meaning and possibly more than one stimulus attributed to an NPC elicitation, the uncertainty can create interest in subsequent sense information with the belief that clarity of meaning will be revealed over time. Ambiguity is accomplished with facial expressions by always allowing multiple simultaneous stimuli sensations and appraisals causing distinct affect selections that can blend in the facial elicitation.

Lastly, *Natural-But-Designed Expressions* are characteristics that value elicitation forms found in nature, such as shapes of the face (eyes, mouth, ears) and their movements that resemble shapes and movements of recognizable natural objects. For example, when eyebrows rise in surprise or rage, they take the shape of volcanoes or other swollen vessels filled with energy. This premise is useful when the avatar is less humanoid or mammalian and more abstract. Consider for example the conversational agent paperclip used as an operating system guide during the 1990s of the Microsoft Windows operating system.

With its flexible deployment, the Affective Loop as applied within a system provides a guide that reinforces an NPCs resemblance to emotionally intelligent behavior and the appearance of living entities. The Affective Loop intends continuous engagement by providing the player with familiar organismic processes.

2.8.3 Subjectively Filtered Imperfect Perception and Memory

NPC emotion models should also allow for occluded, skewed or false perceptions of sensed data by deploying subjectively tuned feature values distorting filters consistent with the NPCs

idiosyncratic personality. Subjectivity in virtual agents is a known topic in software engineering, and it is used to localize agent behaviors to restricted domains, even when the agents are NPCs.[70] Designing subjectivity in an NPC then allows for building "imperfect", non-machine-like or organismic features that simulate human constraints on intelligence. Thus, rationality can be limited as the NPCs character design requires. Similarly, an NPCs memory can also be designed as "imperfect". Memories can be subject to disappearance, erosion, and distortion over time. For an NPC to show signs of organismic entropy is for its memory to simulate decay and self-doubt as the game play proceeds. Furthering organismic limitations, an NPC may not always respond too quickly, nor should it always respond without the delaying force of anticipatory doubt, remorse or regret. Similarly, NPCs can reflect before choosing a course of action. They can likewise elicit a choice with temerity or tentativeness. People often mumble and stumble because they doubt their choices, forget their memories, become distracted by competing but pertinent stimuli, or fail to figure out a course of action in due time. Occasionally unclear elicitation is the result of an overloaded overwhelmed system. Such "imperfections" built into an emotion model make for NPCs that cause curiosity in players and reinforce the lifelikeness of the world that spawned them.

2.8.4 Modularity and Scaling of Emotion Resolution

Less an organismic and more a mechanical concern, the most adaptable emotion models are those that allow for modular components that increase the variety of sensible entities and the resolution of sensing input data. Modularized emotion models can allow an increase or decrease in the number of feature variables an NPC can sense, and an increase in the complexity of elicitation. Modularity gives a designer the ability to use the same model with fewer components for simple mass-produced NPCs, while for more complex supporting characters, and for very

complex antagonists or “buddy” characters, more modules can expand the character’s emotion complexity. Modularity also allows for real time scaling up or down, where if a complex character has a brief and simple appearance, its model can retain less data with fewer components, or vice-versa. Thus, when designing NPC emotion models, scale of entity detection and emotion elicitation may vary enormously, depending on the importance of a character in a game.

Like the limitations of scale, games provide circumscribed interactions with NPCs that reveal an incomplete view of a character. We get to know an NPC over time through a mediated context that serves the whole game. There is no need to model more than what the game design will allow the player to sense during anticipated game play time. A model needs to show only the emotions that the game mechanic allows. Like many forms of humanistic expression that comprise an art form, players must infer what they do not sense. And they cannot sense everything at once.

2.9 Future Research Directions for NPC Modeling

With an emotion model that is modular, it becomes conceivable to implement a procedurally generated emotion as part of a procedurally generated NPC and thus within a procedurally generated narrative system. Procedurally generated game components can become principal features of adaptive games that reshape themselves to the variables of the player. Adaptivity would require some mechanism that can categorically predict the features of the player so that generated game components would spawn in variations that best suit the character designer’s prediction of what the player needs to experience. A predictive system would require a robust machine learning algorithm that enables adaptivity in ways productive to the aesthetic of the game and player-designer relationship. From the perspective of the actor who must perform the character motion, it remains to be seen if the past models presented in this research are consistent with the performance preparation process. That is the work of experimentation to be found

in subsequent chapters. Such experiments intend to address the viability and reproducibility of autonomously animated neural network facial emotion controllers. Their eventual success may lead to a new paradigm for production of photorealistic NPCs and cinematic acting in video games.

2.10 Conclusion

Enabled by processor innovations and the techniques of machine learning, photorealistic NPC design and development may adopt the performative aesthetics of motion picture acting with its complex facial animations. The challenge that lies ahead for developers is to design an emotion model and working process that adapts to existing entertainment methods and techniques. Past practices that relied on hand-crafted morph animations have provided remarkable accomplishments at creating life-like characters for computer games. However, animator-centric NPC emotion modeling based on morph animations is a cost prohibitive method for most game developers and designers. This research promotes a synthesis of the structures of emotion developed by the computational contributions of Cognitive Psychology with the most influential Performance Theories and acting training methods and practice toward the service of procedurally generated character behavior. Future work will reveal if putting a human actor’s creative process at the center of emotion modeling and the workflow of character design will make NPCs more enjoyable and closer to cinematic style acting for computer-based games.

3 RECURRENT NEURAL NETWORKS FOR NPC FACIAL ANIMATION[2]

Creating photorealistic facial animation for game characters is a labor-intensive process that gives authorial primacy to animators. This research presents an experimental autonomous animation controller based on an emotion model that uses a team of embedded recurrent neural networks (RNNs). The design is a novel alternative method that can elevate an actor's contribution to game character design. This research presents the first results of combining a facial emotion neural network model with a workflow that incorporates actor preparation methods and the training of auto-regressive bi-directional RNNs with long short-term memory (LSTM) cells. The predicted emotion vectors triggered by player facial stimuli strongly resemble a performing actor for a game character with accuracies over 80% for targeted emotion labels and show accuracy near or above a high baseline standard.

3.1 Introduction

Cinematic facial animation of game characters has become a requisite ingredient for many emerging video game titles.[71] New technologies of Facial Emotion Recognition (FER) can produce the datasets for game character designers and developers to enable novel techniques and workflows that use machine learning (ML) and NN models. The data produced by FERs can be used to create customized deep learning NN models to control emotion elicitation through the animation of the facial mesh of game characters.

To optimize immersion, game developers rely more on a player's identification with characters through closely placed cameras directed at a character's face modeled from the likeness of renowned actors from films and television. Adding to increasing cinematic realism,

[2] This chapter was published under a different title. Schiffer, S. (2021, December). Game Character Facial Animation Using Actor Video Corpus and Recurrent Neural Networks. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 674-681). IEEE.

the player is seeing less distinction of pixel resolution across media.[72] While it remains to be proven if more complex and nuanced control of facial animation that closely resembles cinematic aesthetics augment player immersion, game researchers are striving to find out.[73] Neural networks as facial animation controllers show some promise to achieve this purpose.[74] But what are the most effective components and parameters in designing neural networks for game characters that autonomously elicit facial emotion expressions? And what is the most effective motion capture workflow for gathering "authentic" emotion elicitation data?

The most frequently used workflows for creating game character animations capture motion data derived from positional transmitters or reflective points placed close to a performer's skin. These markers track commonly located features of the body. Also, motion capture may be gathered from video streams (live or pre-recorded) of an actor performing movements tracked by computer vision algorithms that recognize discernible features of the human body. The motion data is then typically altered by hand-crafted edits applied by animators into the motion controllers of 3d game character meshes. This mixture of computationally generated and hand-corrected animation is labor-intensive and time consuming, but it creates for the player an experience that visually and kinesthetically interprets or simulates human motion of the face, trunk and appendages, resembling a cinematic performance.

This chapter presents a novel approach to developing a facial animation controller and motion data capture workflow that provides a means to elicit appropriate emotion expressions using ML. This experimental research integrates theater and screen character development techniques and rehearsal methods for actors to integrate with emotion recognition technology. The experimental results of this research demonstrate that an effective emotion model can be

developed by combining actor-centric design and neural network architecture for a game character-specific emotion model and development workflow.

The acting teacher and theorist Meisner developed methods for actor training and character rehearsal that resemble the deployment of behavior trees and the training of computational models of emotion using ML algorithms. Some Meisner exercises, when practiced over an extended period, "train" from the actor's affective memory (raw input data) to develop for the character a discrete set of reusable stimuli (predictive output data) for physical elicitations that the actor in-character can use during rehearsal and performance. Meisner techniques train actors to allow the brain to adjust performance by experiencing character sensation and action repetitively with exercises that combine the stochasticity of impulse with immutable the rules of an improvisational exercise.[75] Like a ML model in "training" phase, the actor's character design process focuses and reflects on their stochastic choices to calibrate their character "weights and biases" for the next performance. The process of an actor calibrating through repetition resembles the corrective process of ML, where the weights and biases of a neural network are altered with every iteration of a Back-Propagation Algorithm Through Time. In our experiment that we will describe below, the use of a multi-layered RNN with Bi-Directional Long and Short-Term Memory (BD-LSTM) cells combined with an output layer of Time Distributed Dense perceptrons provide a comparable model to some Meisner exercises.

Previous approaches to automated facial animation controllers have derived emotion data from general facial emotion datasets but without data from actor-elicited emotion.[76,77,78,79] This research proposes creating a specific corpus of facial emotion video developed by a single actor for a single character. We found no other research that has taken this approach. The objective is to demonstrate a novel emotion capture workflow and ML algorithm for game character facial

animation that resembles the actor's process of training and preparation. The resulting facial motion data can imbue a game character with the emotional intelligence of a specific actor's interpretation.

3.2 Related Work

Past work that informs this research requires an interdisciplinary approach that draws from acting (performance) theory and ML techniques. As bridging disciplines between these two, computational cognitive psychology and affective computing provide important contributions as they both model and quantify emotion elicitation and propose computational models of emotion that represent and simulate human emotion systems for synthetic agents like game characters. Related work in these fields each tackle distinct angles to the problem of developing autonomous emotion elicitation systems for animated characters in video games. This chapter will summarize the most significant contributions from specialized areas that have enabled the experiments of our research.

3.2.1 Psychological Models of Emotion

Computational cognitive psychology evolved appraisal theory to explain how human emotions are derived and structured in the mind, and it developed a theoretical and practical basis to measure them. The objective was to develop a theory that could be translated to software. Ortony, Clore & Collins (OCC) describe emotion as the result of "appraisal" occurring as a person assesses their environment and its impact on affirming standard and tastes, choosing goals, and determining the action needed to attain them.[80] The OCC framework continues to influence the modeling of emotion AI for NPCs[81] and virtual agents.[82,83] Contemporaries in Cognitive Psychology, Russell, Mehrabian and Scherer among others, evolved a system that quantifies pleasure, arousal and dominance with circumplex models that inscribe emotion as

normalized vectors mapped within a unit circle or sphere [9][10].[84,85] These conceptual models for emotion give impetus for this research to implement normalized emotion values as linearly interpreted vectors deployable in "blendshape" animations within a game engine.

3.2.2 Models of Emotion in Acting and Performance

A bit less ordered are the theories of emotion provided by performance theorists and practitioners. It is alleged that Stanislavsky evolved from French psychologist Ribot, the term of art, "Affective Memory,".[86] The concept of creating a character with human affective memory through repeated action exercises, draws from the Meisner techniques of character development. Clurman advanced a technique of scoring memorable affect with goal-action focused performance annotation[87] that provided generalizable constructs for computational implementation. Combining Meisner's repetition exercises[88] with Clurman's character analysis schema, connects the preparation of actors with that of software engineers by organizing the set of character actions as a performable dataset that this research has implemented as an algorithm and set combinatorics in a graph.

3.2.3 Computational Models of Emotion

Informed by models of emotion from theorists in psychology and theater, academic software developers of video game characters and virtual agents designed systems for synthetic emotion elicitation. The Oz Project, a collection of video game experiments and research papers realized by Loyall, Bates and Reilly, made extensive use of emotion generation processes for actors in training.[89] The Method, as developed by Stanislavsky and advanced and revised by Meisner and Clurman, and the emotion system structure of OCC, were implemented in the digital artifacts of the Oz project. Loyall et. al produced several interactive media experiments with autonomously animated digital characters using emotion AI systems based on theories from psychology and

theater.[90] More recently, Kapoor considers the unique problems of designing and evaluating the ML models for affective computing, specifically the problem of creating valid emotion labels, adapting to the variability of quality and quantity of affect databases, and overcoming the difficulty of defining what is a ground truth for comparing an emotion prediction with its "true" referent.[91]

3.2.4 Commercial Game Engines

In 2019, Unity3d announced its ML-Agents package that enabled developers to either create NN models within Unity or to compile them as externally linked models. Likewise, in 2020, Unreal Engine released its Python Plugin to facilitate externally linked NN models. Facial animations require forms of AI and ML to provide time-series predictions frame-by-frame of facial muscle systems. But no system in this research was found to have implemented a NN model trained from a specific actor's video corpus and dataset for facial emotion animation of game a character. Using a FER system as a tool for actors to create emotion data for game characters, this research deploys supervised learning on the emotion vectors in time-series datasets. The resulting NN models create predicted emotion vector outputs that animate groups of nodes of a 3d facial mesh within the game engine.

3.2.5 Facial Emotion Video Corpora and Datasets

New markets for video corpora libraries are expanding for use in governments and industries. However, questions arise concerning emotion recognition accuracy. Two problems are: (1) Not all emotion labels are equally represented in video corpora. *Disgust* for example, is hardly found in many corpora. *Surprise* is also underrepresented.[92] Video corpora with unbalanced emotion representation provide less data from which neural networks can train, thus causing biases toward the more densely represented emotion labels. Panda cross-validated some of the most

widely used “in-the-wild” video corpora and their datasets of human emotion and found 22.99% misidentification of negative emotions.[93] Misidentification was frequently correlated with image features unrelated to the face, thus undermining the reliability of “in-the-wild” video corpora. This research elected to model its corpora production on methods used in the development of acted dyadic corpora, principally the MSP-IMPROV corpus, with its emphasis on fixed lexical content with varied emotion targets for elicitation,[94] and the CreativeIT database, for its use of verb-defining actions to elicit goals through acquisition challenges.[95]

Creating studio-produced video corpora used specifically for game applications, as this research proposes, is an actor-centric alternative. Most video corpora and datasets consist of images using a moderate to large number of subjects performing a variety of tasks. In preparation for this experiment, no video corpora were found to consist of hours with the same subject performing a fictional scenario with a determined set of variations. Extent video corpora intend to provide data that can be generalizable to classify faces outside the set of subjects in the corpora. The opposite, where the subject’s corpus intends to be used to predict a model behavior of that same subject, has scarcely been attempted. Aside from the entertainment industry, where actor characterization is a driving attraction for an audio-visual product, there would be little incentive to create such a corpus. But Busso et.al showed that using actors to create corpora that target specific emotion elicitations can provide more balanced datasets with less bias.[96] With the rigorous and optimized workflow that this research proposes, the benefit of elicitation accuracy provided from targeting specific emotion labels makes actor-centric corpora more feasible. Reliance on trained actors who can perform “authentic” emotion elicitation in an artificial environment becomes tantamount to simulating “natural” elicitations typically found in many video corpora and their datasets.

3.2.6 Facial Emotion Recognition (FER)

The biases of FER systems can be partially mitigated by using those systems that allow calibration for each face. Noldus FaceReader 8 is the FER system used in the corpus of this research, and FaceReader 8 provides a calibration tool adjust emotion recognition to each subject's face. The recognition techniques of FaceReader 8 rely on convolutional neural networks and have resulted in the product's high ranking for accuracy in comparison to competitors[97] and has been externally validated with 88% accuracy on the Warsaw Set of Emotional Facial Expressions (WSEFP) and the Amsterdam Dynamic Facial Expression Set (ADFES).[98] Noldus has disclosed that the video corpus used to train the FaceReader emotion recognition NN is the ADFES.

3.2.7 In-Game Neural Networks

Kozasa et. al showed an early use of an affective model for an emotive facial system in an NPC based on an invented dataset of expressions.[99] Theirs used a 3-layer feedforward artificial neural network to train an NPC from invented data for parameters fed to a NN model as they claimed no databases at the time existed to train their model. Later, using appraisal-based design from virtual agents, the FAtiMA architecture was integrated with a NN model in educational games.[100] Its implementation in its earlier versions in social and educational games has proven to be effective for learning and engagement.[101] The use of LSTM cells for emotion recognition of facial video was shown to improve previous NN performance.[102] But unlike the methods that this research proposes, these previous works did not use NN models whose emotion recognition training is drawn from single-actor video corpora generated for game character roles.

3.3 Methodology

The experiment design involves three stages: (A) *Designing a Dyadic Behavior-Dialog Graph.* Two human characters engaged in conversation follow each of the paths of an acyclic graph with multiple paths from a single starting node to a different single ending node. (B) *Actor-Character Video Corpus and Dataset Design.* A video corpus is recorded consisting of video shots of an actor performing all possible paths through the acyclic behavior-dialog graph. A FER system classifies and quantifies the data to create a meta-dataset. (C) *Emotion Model* Design. A NN model is trained and tested from the FER-derived emotion measurements to implement within a game engine for player interaction.

3.3.1 Designing a Dyadic Behavior-Dialog Graph

Scenes between a player character and an NPC often require an exchange of words and gestures. Game input sources can use the camera trained on a player's face to recognize emotions and trigger NPC facial animation. But for an NPC to respond autonomously and appropriately to a player's facial elicitation, its animation controller must be trained to recognize and interpret emotions from a player's face. This research proposes that an actor's performance playing the role of an NPC be used as the basis for training a NN that controls a game character's facial animation. Following an appraisal model of emotion, the NPC seeks to affirm its *standards* and *tastes,* but more importantly assess a player's utility toward achieving the NPC's *goals,* and to select a responsive NPC *action* toward the player. To prepare for the dyadic emotion elicitation recordings between a "stand-in" player and an NPC, actors must be briefed on the player's and NPC's beliefs, goals and actions the character could take to get their goals and remove obstacles. This research demonstrates that an actor's character design demonstrated through facial expression performance can be modeled from automatic FER data into a NN

model that predicts a reaction to the player character using emotion label values, and thereby a facial expression that is probabilistically in the range of the actor's choices. The prediction becomes the data to direct an animation controller of an NPC's face. This research does not demonstrate rational decision processes, but only emotion elicitation in response to player expressions. The experiment focuses on the viability of the NN architecture, but to discern viability, a structured corpus had to be created.

The first assumption is that a collaborating actor can elicit an authentic and recognizable targeted emotion on camera for a FER system to measure. To create intended elicitations, the experiment used a six-node acyclic behavior-dialog graph with a dyadic scenario that could be resolved between two characters using minimal scripted dialog. A pre-recorded player character called the Stimulus Source presses with straightforward accusatory questions for information in a police interrogation. With the Stimulus Source playing detective, they seek a confession to trigger emotions from an on-camera character called the Model Subject, also played by an actor. Actors for both characters performed asynchronously in chairs facing a camera. The Stimulus Source used a few short sentences that the Model Subject must respond to with the words "Yes", "No" or "So", or some short-improvised variant of those three words. The eight-nodes provided 32 possible paths with one starting node distinct from one ending node. An abstract representation of the graph is shown in Figure 3.1.

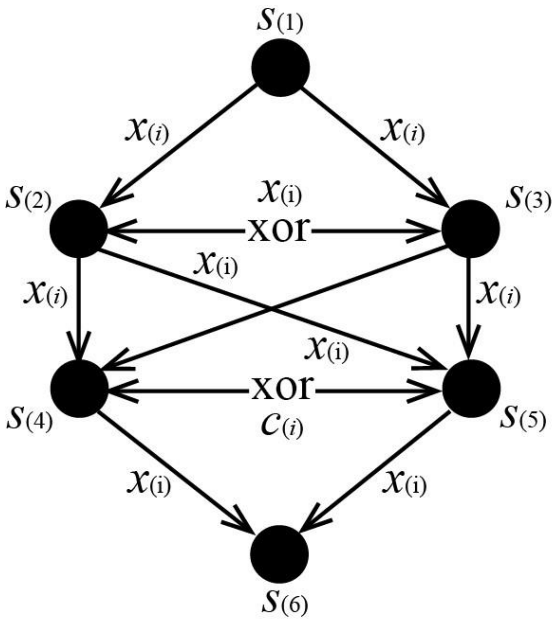

Figure 3.1. Behavior-Dialog Acyclic Graph. For emotional states $S = \{s_1 \ldots s_6\}$ and actions $C = \{c_i \ldots c_n\}$.

What follows is an abbreviated description of the rehearsal and production method for creating the video corpus and dataset. We use graph algorithm terms for efficient description for reproduction of the experiment. We can let graph $G = \{A, S, X, P, E\}$ so that G consists of sets S of states, X of actions, A of attributes, P of paths and E of emotions. Movement through the graph is recorded by the whole dataset T of 12-dimensional vectors each at time t_i formed of two concatenated emotion vectors. The pair are emotion vectors representing synchronized video frames analyzed by the same FER system. The frames analyzed for one in the pair are of video shots of the Stimulus Source and the other the Model Subject. Each are performing one of 32 paths p_i of P of the behavior-dialog graph G, and the result of frame analyses from each of 288 video shots of the Model Subject. The two concatenated vectors contain 6 normalized values of emotion labels (*anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*), where each t_i represents a fraction of a second of the recording of the faces of the actors.

A is a set of scenario attribute sets encountered during every path p_i of *P* through graph *G*. Let $A = \{L, O, H, C\}$ where *L* is a set of goals l_i that the Model Subject pursues in the scenario and in each path p_i of *P*. *O* is the set of obstacles that obstruct the Model Subject from getting any goal l_i of *L*. H is a set of thresholds h_i affiliated with each emotion label e_{ii} of E_i. And *C* is a set of actions available to the Model Subject to remain in state s_i or to move along edge x_i and on to the next state s_j. Each state s_i of *S* consists of a series from *T* of timesteps t_i where one of the values e_{ii} of emotion label E_i of *E* exceeds in value beyond threshold h_i of *H*. Each state s_i allows the Model Subject to react and decide toward the Stimulus Model to attain goal l_i of *L*, or to continue to respond to the character's internal (unspoken) thoughts and imaginations.

X is a set of edges where x_i is an action from a set of verbs chosen by the director of the actor to get a character's goal, l_i, or to circumvent or remove obstacle o_i of *O*. The edges of *X* are used to move from one state s_i to the next s_j across a series of timesteps t_i in *T*. Each edge x_i allows the Model Subject to act toward the Stimulus Model to circumvent or remove obstacle o_i from one of 32 paths p_i of *P*.

P is a set of unique paths p_i through the acyclic behavior dialog graph such that any p_i consists of a combination of action edges x_i of *X* and state nodes s_i of *S* where the first in the combination is always start node state s_1 and the last is s_6 of *S*. E is a set of emotions where $E = \{E_i \ldots E_n\}$ for prediction of the NN from the resulting video corpora, and $E_{i \ldots j}$ are each of the emotion label columns in the set. E_i contains the set of all column values $e_{ii} \ldots e_{in}$ where the first subscript *i* affiliates with one of the six emotion E_i and the second subscript *i* is the index of the tuple row. Subscript *n* is the total number of tuples in the dataset *T*. Since every timestep t_i within a state or action is a tuple in the dataset *T*, all values $e_{ii} \ldots e_{in}$ of E_i have one value in every tuple t_i. When the value of each of the emotion labels E_i equals or exceeds a threshold h_i of *H*

for each emotion label E_i of E, then the condition for the Model Subject exists to react outwardly toward the Stimulus Source to the next state s_i of S as an action on any edge x_i of X. An action x_i is a conscious choice that will trigger a facial animation realized as a change in values $e_i \ldots e_j$. An algorithm that describes the Model Subject's progression through the 32 paths p_i of P in the scenario follows in Listing 3.1.

LISTING 3.1: ALGORITHM FOR EMOTION MODEL PROGRESSION THROUGH DIALOG-BEHAVIOR TREE

Input: vector t_i of T.
Output: Some action x of X

1. for $t_i = 1$: n do
2. if ∀ values e_i of t_i >= threshold h_i of H
3. select next action x_j from X
4. return x_j
5. else
6. return repeat x_i

3.3.2 Actor-Character Video Corpus and Dataset Design

A camera recorded the Stimulus Source performing all possible paths p_i of P. All nodes and edges were covered without traversing any edges more than once. Using those recordings, precisely the same video segments of nodes and edges were copied and edited to construct the 32 videos of paths p_i of P of the Stimulus Source performing their side of the dialog for the Model Subject to react. While watching the Stimulus Source perform 32 videos paths of P, the Model Subject performed 9 video recordings for each of the 32 video combinations. The 9 video recordings were presorted into triplets where the actor was given different narrative information that would increase their arousal such that the first 3 takes would illustrate the lowest arousal, the next three takes increased arousal to a mid-range and the last 3 takes would present the maximum arousal. The Model Subject was also directed with narrative information to play for goals in the scene that would likely elicit 3 targeted emotions (*anger, sadness* and

fear) of the character to best shape the player's perception and identity of the character portrayed.

The total number of video recordings created from which to eventually extract facial emotion expression data with a FER was set as the product of the number of video shots of each path (9) times the number of possible paths (32) through the dialog behavior tree. For this experiment, that total was 288. The scenario and all the traverses through the dialog-behavior graph were rehearsed in advance and designed such that all recorded performances of the Model Subject would be at most 60 seconds. The experiment set its video recordings at 24 frames each second. From the 288 video recordings of approximately 60 seconds each, there were 414,720 frames of video for the FER systems to analyze. The FER system was set to analyze 1 frame out of every 8, thus there were 3 emotion analyses performed each second. The FER systems generated 51,840 tuples from the 288 recordings of the Model Subject and the Stimulus Source.

3.3.3 Emotion Model Design

One stated goal of this research is to design a neural network architecture that would reflect the essential stages of observation and creation experienced by actors in preparing for a role. An actor is still a human first, and as a human the actor already comes with sentient awareness, standards, tastes, goals and behavioral filters. When an actor is cast and before they begin preparing their role, the emotion system of the character and the actor behave the same. But as an actor begins to perform the required actions described in the character's scenario, a split occurs where the actor creates a subset of sensitivities that belong to the character. This subset evolves as an actor's emotion system, with its "weights and biases", amplify or increment the character's stimuli through preparation and rehearsal. Actor preparation is a process of recalibrating the

emotion elicitation system for a character and its context, and at times differentiating one's own emotions from those of a character.

Following Meisner's methods, an actor maintains sentient awareness through the filter of the character, while allowing the emotional potency of the stimuli of the past to grow or diminish over time. Thus, the initial model of this research consists of a first layer of 100 LSTM cells. This model would provide for training with a maximum residual memory of 100 tuples each representing 1/3 of a second, or the equivalent of up to 33.33 seconds of the past relative to the position of the tuple where the training algorithm is processing. For this experiment, the NN timestep length variable was set to 30 to allow backpropagation to be influenced by values stored from as far back as 10 seconds from the currently processed tuple. The model uses the timestep value to control how much the model allows the recent past to influence emotion predictions in the immediate present.

Additionally, an actor eventually memorizes the scenario after many rehearsals and long before completing 288 similar recordings. While he or she may be sensing "in the moment" as Meisner-trained actors are encouraged to do, in truth a part of the actor's and character's consciousness can and does anticipate the future. Thus, the neural network of this research incorporates a bidirectional layer of LSTM cells, as was used in,[103] that allows the values held in them to represent those from the future. These two oppositely directed LSTM layers are of equal length (100). Their connecting NN edges are processed together with a conduit layer called a Repeat Vector for as many as the value of timesteps; again, in this case, the value is 30. Finally, the output is a layer of Dense connected perceptrons whose output consist of only 1 value.

The choice of one output value means that at most, the model can only predict one of the emotion labels values e_{ii} of E_i. Thus, the emotion elicitation described herein trains a prediction

team using 6 separate NNs, one for each emotion label, each predicting only the value for one of e_{ii} at each timestep t_i. To animate a facial 3d mesh using the methods proposed, all 6 neural networks must simultaneously plug into the 3d mesh model; each must respond to data from timesteps produced 10 seconds in the past and 10 seconds into the future, if future data is available. Training the 6 neural networks used the same parameter setting. 211 batches of 30 timesteps (forwards and backwards) were divided into 6 samples. Each NN model was tested to find an optimum epoch count following two recursive progressions. The first progression of increasing epochs used a rate $n_j = 2n_i$ with a range [10, 640]. The second progression of increasing epochs used a rate of $n_j = ni^2$ with a range [10, 1000].

3.4 Results

To determine which of the trained NNs are the most accurate for each of the emotion labels, the experiment validates that the targeted emotions were sufficiently recognized in proportion to a set of value thresholds (0.01, 0.05, 0.10, 0.20, 0.30) as shown in Figure 3.2. The accuracy of all generated predictions is differentiated by epoch count in relation to two degrees of precision (0.1, 0.01). Accuracy is defined in five ways. (1) A comparison is made between the predictions of each NN model with a baseline prediction as seen in Figure 3.3. For this experiment, the baseline prediction was calculated as the median value of ground truth values in a moving window for t+30 timesteps in the future and t-30 in the past from timestep t. (2) A calculation was made of the percentage of predictions whose predicted value in relation to the ground truth value is above the baseline with both degrees of precision as shown Figure 3.2 and Table 3.1. (3) A computation of the mean squared error between predicted values and ground truth values as seen in Table 3.1. (4) A computation of the standard deviations of the accuracy percentage as shown in Table 3.1. (5) A comparison of the size of the range of values across 23 tests as shown in

Figure 3.2 and Table 3.1. The scenario used to produce the video corpora gives data for all 6 emotion labels, but anger, sadness and fear were targeted in the scenario using Clurman's script scoring technique to evoke these emotions from the actor.

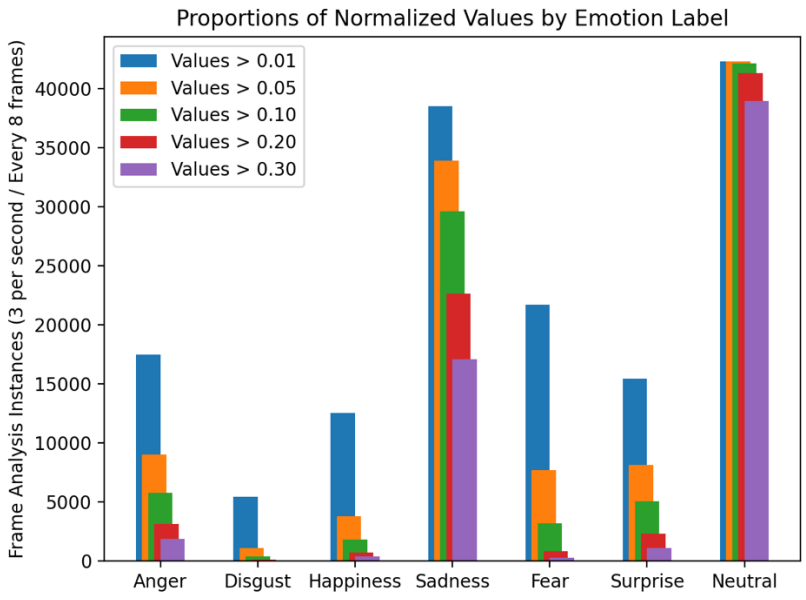

Figure 3.2: Proportions of Normalized Values. Over ~45,000 frame instances.

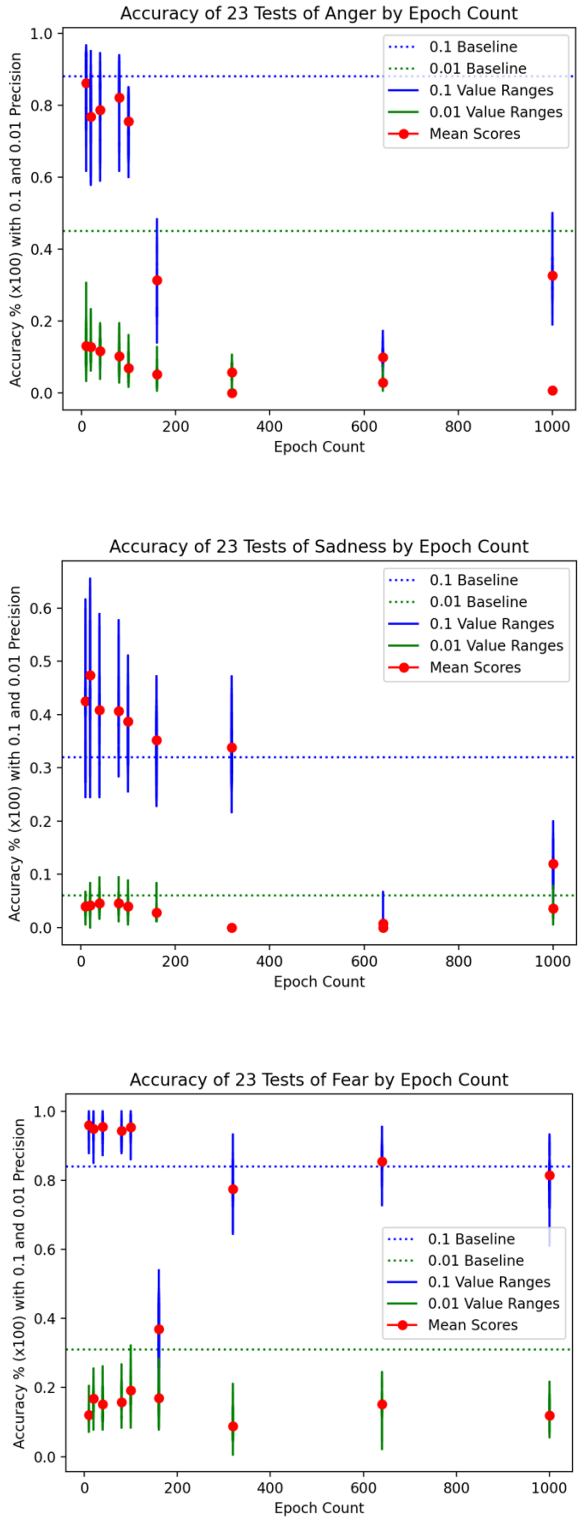

Figure 3.3: Targeted Emotions Anger, Sadness and Fear. NN performance indicated by magnitudes of accuracy. Anger and Fear are most accurate.

TABLE 3.1: OPTIMAL MODEL ACCURACY BY EMOTION LABEL

Emotion	***Precision***	***Epochs***	***Stand. Dev.***	***Range***	***MSE***
Anger	0.01	10	0.0678	0.2722	0.0129
Disgust	0.1	10	0.0059	0.0222	0.0004
Happiness	0.01	10	0.0831	0.2889	0.0033
Sadness	0.1	10	0.1068	0.3722	0.0740
Fear	0.01	10	0.0325	0.1333	0.0026
Surprise	0.1	10	0.0542	0.1889	0.0070
Neutral	0.1	10	0.0713	0.3333	0.0539

3.5 Discussion

One objective of this experiment was to find an NN architecture that could perform objectively better than the baseline. The baseline as defined above is a linearly interpreted smoothing algorithm for data that is already relatively smooth. For a NN model to approach or outperform a baseline that is a smooth shadow of the already smooth ground truth would suggest that the NN model could be a productive solution for autonomous animation. The proximity of all accuracy tests with 0.1 precision suggests that if put into a production workflow, the methodology will provide game character facial emotion elicitations that substantially resemble the actor who performed in the video corpus. Another objective was to find an optimal epoch count for each of the emotion labels. The results show that lower epoch counts will produce more successful results in every test, even for those emotions that were not targeted in the scenario. This fact is substantially important since the elapsed time to train an NN on 10 epochs is two full degrees of magnitude less.

A subsequent experiment should consider the effect of varying the number of timesteps. 30 timesteps or the equivalent of 10 seconds were used, and these represent approximately 17% of each video recording. The timestep length was determined by the pacing of the performance of the scene. The exchange of a few words and gestures between the Model Subject and Stimulus

Source was rapid enough to assume that after 10 seconds, the emotional impact of whatever occurred was supplanted by new stimuli. If the timesteps were too long, then the model's memory may be longer than the characters. If too short, then the model misses the effects of influences of behavior the performer experienced.

Another point for further consideration is the reliance on the actors for pacing their character through the behavior-dialog graph. They are not able to see if their elicitations of any one emotion label exceed a threshold h_i of H to allow progression to the next node. The methodology section provides an algorithm that suggests that such a threshold exists. Actors are trained to mark in their memory the threshold of emotional responsiveness necessary to push them to take new action. But this threshold exists in the mind of the actor and is reinforced in the character (within the actor's mind) through rehearsal, without a physically attached observation device. Calibrations in the actor's performance depend on subjective observations that are adjusted by oral requests from the performance director between recordings. Regulating an actor in the middle of a performance often skews the emotion analysis results. This limitation remains immutable, for regulating an actor in mid-performance would disturb the actor's sentient focus on the fictional reality constructed in the imagination and would interrupt the emotion elicitation.

A discovered correlation is that the more frequently an emotion label had values closer to 1.0, the worse the model performed for accuracy. All models for each emotion provided values between 0.0 and 1.0. Both neutral and sadness have greater frequency of higher values. But their emotion models performed with the worst accuracy. Conversely, Disgust and happiness showed a greater frequency for lower values, but their accuracy measurements were among the best. Meanwhile, surprise and fear showed middling values in relation to the other emotion labels, but

their accuracies remained relatively high. Proportions of emotion label values are reported in Figure 3.2.

3.6 Conclusion

While the experiment suggests the methodology proposed is viable, further experiments need to determine which parameters and conditions are mandatory for the desired result and to optimize the most expensive steps. The next experiments will address the following. (1) The size of the behavior-dialog graph. Adding more node levels will linearly increase the number of video shots required to cover the graph. Increasing the graph size and thereby increasing the number of elicitation cycles will require more video shots. Will this increase in video corpus size improve the accuracy of the model? (2) The value of the training parameters. While a lower number of epochs showed optimal accuracy results, can accuracy improve by adjusting the number of batches, the number of timesteps, the number of samples per batch? (3) The NN Component Count. The number of LSTM cells in the two layers were fixed. Will decreasing or increasing the LSTM cell count improve or worsen accuracy? (4) Implementation of teams of NNs. Team implementation needs experimentation within a fully realized game engine that controls the complete facial animation of a game character such that a FER can analyze the emotion elicitation of the animated NPC and compare its elicitation classifications and measurements with original actor video shots.

3.7 Future Work

Creating an actor-specific video corpus for each major game character would be feasible if each game character was extensively used in a game and reused in subsequent editions of a game franchise. The experiment showed that deploying a FER system that uses CNNs and allows for elicitation calibration of individual actor faces, can provide sufficiently accurate data to produce

NNs that can predict sub-second instances of three or more facial emotion expressions. Creating an actor-specific video corpus trained for 10 epochs using 211 batches on a bi-directional auto-regressive neural network consisting of a forward-directed neural network layer and a backward directed layer of 100 LSTM cells each, will predict very similar emotion data with 0.1 precision. Animation of photorealistic human faces of game characters is subject to high standards of resemblance to the movement of real human faces. Game players expect character faces to move fluidly from one emotional elicitation to the next in response to the stimuli of the character's environment. If the animation quality driven by a team of neural network models provides sufficient resemblance to that of motion pictures, then the experiment may prove that a new workflow for game character animation may be feasible for general implementation.

4 FACIAL EMOTION EXPRESSION CORPORA FOR TRAINING NPC NEURAL NETWORK ANIMATION CONTROLLERS[3]

While numerous video corpora have been developed to study emotion elicitation of the face from which to test theoretical models and train NNs to recognize emotion, developing single-actor corpora to train NNs of NPCs in video games is uncommon. A class of FER products has enabled production of single-actor video corpora that use automatic emotion analysis. This chapter 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.

4.1 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 behavior 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 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.

[3] This chapter was published under a different title. 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 – Volume 2, SciTePress, pages 197-208. DOI: 10.5220/0010874700003124.

While agencies in government, military, security, and product marketing have a great interest in predicting the thoughts, emotions, and behaviors of large sets of randomly selected “real” people, the entertainment industries make their business by creating and predicting the thoughts and behaviors 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 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 skillful 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.

4.2 Related Work

The experimental design of this chapter 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 chapter will summarize findings related to acted corpora and some "natural" corpora where "real" people are used as behavioral subjects. Additionally, deployable acting theories for corpus production will be briefly 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.[104] Additionally, in the case that actors performed as characters, were the elicitations induced using another technique known as (personal emotion memory) substitution,[105] or was the emotion produced from a dyadic conversation using improvisation principles of active analysis technique.[106] As this chapter 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 chapter and its experiment advances, it is important to address the types of corpora production which the experiment of this chapter 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 chapter and its experiment considered if the on-screen subjects were aware of the camera's presence or if they were 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[107] largely through what has been called a neurologically "constructed emotion".[108] Furthermore, the affective computing community has remarked on the ethical concerns[109] and the accuracy of using such footage for corpora development to study or model emotion due to inhibition of emotion display.[110] 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.[111] Our research concludes that the corpus sample set to use for modeling 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 modeling). The input emotion should categorically resemble that of the predicted output.

Research for this chapter 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.[112] 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".[113] 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."[114]

The Interactive Emotional Dyadic Motion Capture database (IEMOCAP) 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.[115]

The Multimodal Signal Processing Improvisation (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.[116] 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 the 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 Meisner, a teacher and theorist. The technique is widely known as the Repetition Exercise in acting communities and the exercise has many variations still practiced by Meisner technique instructors.[117]

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 chapter 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 corpus shared one goal of this chapter and its experiments. It was the only corpora that explicitly set out to create a corpus that would assist in analyzing "theatrical performance" for entertainment in addition to the broader application of "human communication".[118]

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.[119] 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 analyzing context can be applied to acted corpus sample production provided the sample and the rehearsal have sufficient duration.

Our research found useful the MSP-Face Corpus,[120] another of the corpora consisting of video-sharing content. Its production methods were completely different from the experiment of this chapter. 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.

4.3 Methodology

The controversy over using actors to create samples or to collect samples of "real" subjects found "in the wild" 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.

4.3.1 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 behavior while also concentrating on the fictional stimulus that triggers it.[121] The performance 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 performance director to set up 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.[111] 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,[122] or a team of NNs can be trained to animate each specific emotion.[123] 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.

4.3.2 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.[124] Corpus validation for acted elicitation has shown success when actors were allowed to prepare in the manner they were trained.[125,126]

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 chapter. 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.[116]

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 chapter. 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,[127] 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 4.1 presents one of the actors in-character during production of a single-actor corpora in a studio environment with timecode slate. Figure 4.2 illustrates the setup of the recording of each sample.

Figure 4.1: Corpus Actor Alfonso Mann.

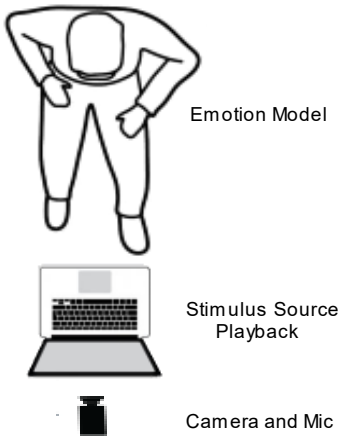

Figure 4.2: Production Setup. For sample recording.

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.[128,129] The experiment of this chapter deployed the technique by designing a dialog-behavior 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-behavior tree shown in Figure 4.3 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-behavior 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.

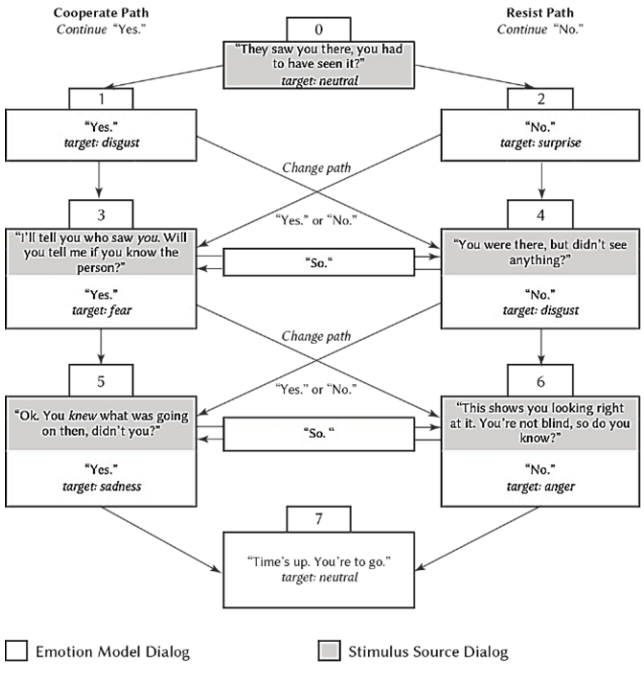

Figure 4.3: The 8-Node Directed Acyclic Graph. Produces 32 unique video clip variations. No nodes or edges repeat. Grey dialog was performed by the Stimulus Source. White dialog was performed by the Emotion Model.

Using a Depth-First-Search algorithm,[130] 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[131] such that for each dialog or behavioral "beat" (which in this case is represented as a node), there is one tactile goal represented by a noun (an objective), and one behavior 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.

4.3.3 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 behavior, 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 behavior of NPCs. Thus, when players observe an NPC's behavior 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 behaviors. Rapidly players notice behavior 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-behavior tree.

4.3.4 Design of Emotion Elicitation Content: The Game Scenario Dialog-Behavior Graph

An examination of the structure of the dialog-behavior graph reveals that, like any tree graph, it has a height and width dimension. In the use presented for the experiment of this chapter, 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.

4.3.5 Sample Recording and Data Generation

The OMG-Emotion dataset emphasized the importance of corpus sample production methods that provide contextual data around facial elicitation.[132] 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 chapter. 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 behavior 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.

4.4 Results

Unlike the corpora reviewed for this chapter, 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 analyzed 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 4.4 is edge D. Figure 4.4 shows a node-to-edge version of the dialog-behavior 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.

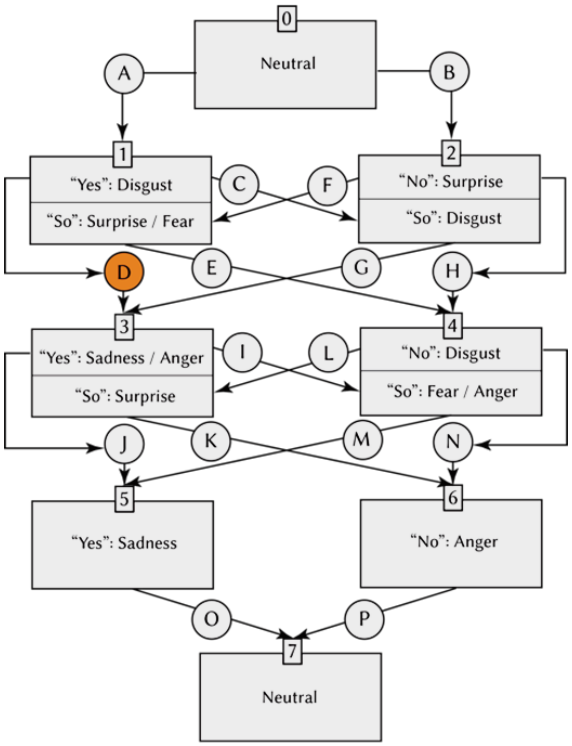

Figure 4.4: Dialog-Behavior Graph for Dyadic Interaction. Edges (A to P), nodes (0 to 7). Model responses in quotes. Orange node-D is the focus of this study.

Table 4.1 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 4.1 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 4.4, 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 4.1: INSTANCES OF EDGE D IN EACH SEQUENCE

		…	Seq. 5	Seq. 6	…	Seq. 9	Seq. 10	⋯	Seq. 21	Seq. 22	…	Seq. 25	Seq. 26	…	Sum of Instances
Nodes	Timecode		013457	01357		0213457	021357		013467	0136		0213467	021367	…	
1	0:14.6	…	1	1	…	1	1	…	1	1	…	1	1	…	8
3	0:26.6	…	1	1	…	1	1	…	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.

4.4.1 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 4.5 and 4.6 show. Statistical analysis over all the values produced by frames analyzed by the FER system in the edge of each path in the dialog-behavior 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.

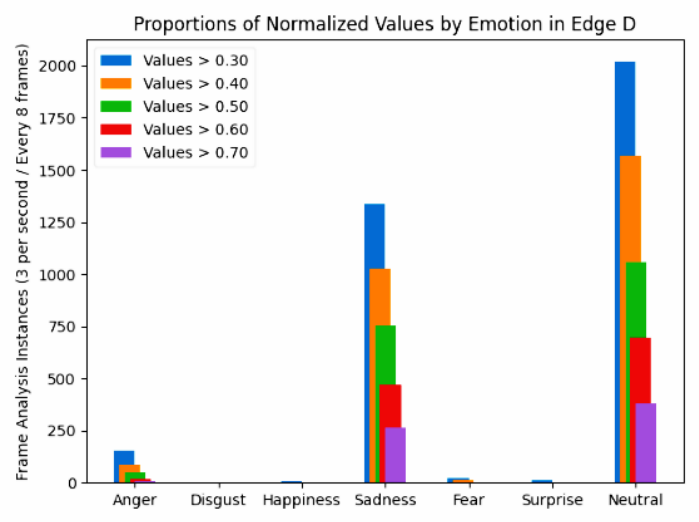

Figure 4.5: Performance of Edge D. Reveals sadness is primary emotion, followed by anger.

Figure 4.4 identifies the position of edge D in the dialog-behavior graph as an edge leading into node 3. The primary emotion is sadness, and the secondary emotion is anger. Figure 4.5 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 4.6 and 4.7 show proportions of emotions sadness and anger over the same segment of timecode. Figure 4.8 confirms the higher levels of sadness lingering with a gradual decline that gives way to anger for a climactic reaction as shown in Figure 4.9.

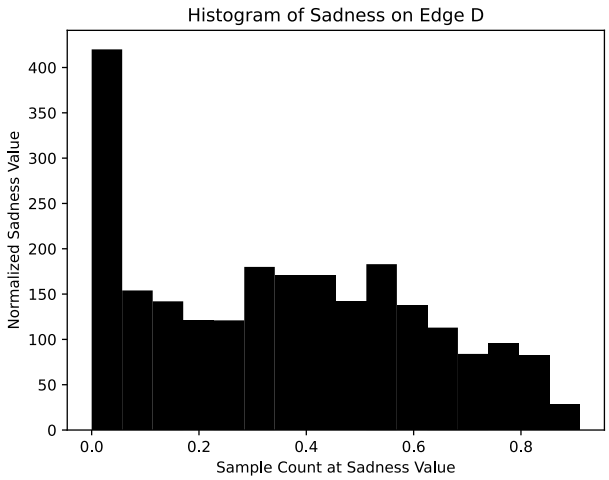

Figure 4.6: Histogram of Primary Emotion Sadness for Edge D.
Indicates maximum is 0.910 and the median is .384

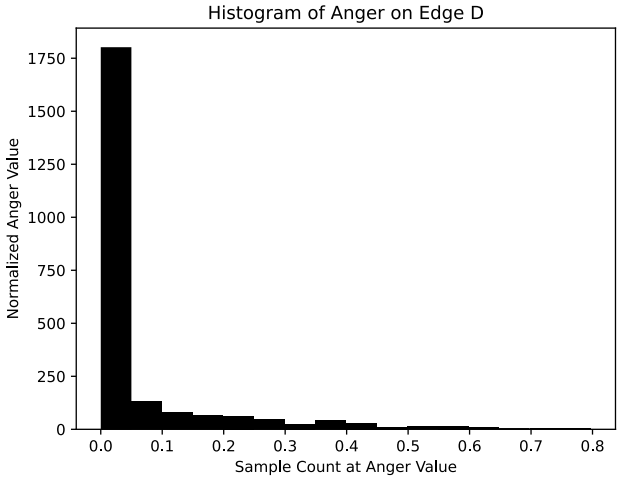

Figure 4.7: Histogram of Secondary Emotion Anger for Edge D.
Indicates maximum is 0.797 and the median is 0.0000006

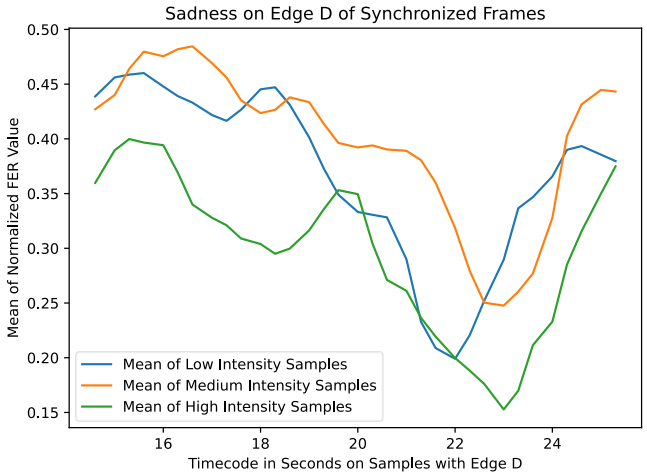

Figure 4.8: Primary Emotion Means of Sadness on Edge D for Three Intensities.

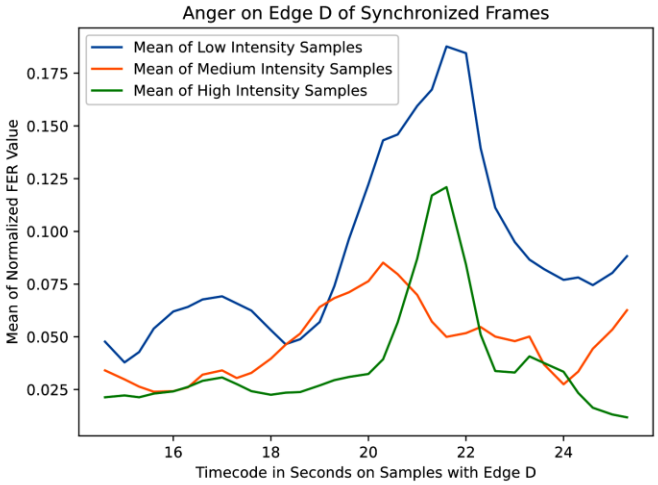

Figure 4.9: Secondary Emotion Means of Anger on Edge D for Three Intensities.

It should be noted that the intention of the intensity triplets was not adhered to. 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 4.4 for edge D was elicited in the segment intended.

4.5 Discussion

This chapter 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-behavior tree with the intention of producing emotional variability can prevent overfitting. The dialog-behavior 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-behavior tree, unsupervised learning can train a NN without overfitting. But will optimization best be served by adjusting only the width of the dialog-behavior 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-behavior 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 analyze 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.

4.6 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.

5 MEASURING EMOTION INTENSITY FOR EVALUATING RESEMBLANCE[4]

Game developers must increasingly consider the degree to which animation emulates the realistic facial expressions found in cinema. Neural network controllers have shown promise toward autonomous animation that does not rely on pre-captured movement. Previous work in Computer Graphics and Affective Computing has shown the efficacy of deploying emotion AI in neural networks to animate the faces of autonomous agents. However, a method of evaluating resemblance of neural network behavior in relation to a live-action human referent has yet to be developed. This chapter proposes a combination of statistical methods to evaluate the behavioral resemblance of a neural network animation controller and the single-actor facial emotion corpora used to train it.

5.1 Introduction

As expensive as they are to design and produce, photo-realistic human agents have become a common attraction in contemporary video game design of non-player characters (NPCs). To get them to behave with emotional veracity, video game developers are using AI techniques to control facial expressions. Developers may choose between at least two approaches. The first evolved through the Computer Graphics research community. It prioritizes mimetic resemblance of movement and modelling to the appearance of a performing actor or model. The second was developed by the Affective Computing community, and prioritizes emotional resemblance, which is the ability of the avatar to autonomously elicit a facial expression based on an integrated emotion model. These two approaches have evolved over several decades using different models

[4] This chapter was published under a different title. Schiffer, S. (2023). Measuring Emotion Intensity: Evaluating Resemblance in Neural Network Facial Animation Controllers and Facial Emotion Corpora. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SciTePress, pages 160-168. DOI: 10.5220/0011655300003417

of simulation. The former creates a system of expression generation based on appearances on the surface of the agent's face. The latter attempts to encapsulate an emotion generation system that is located "inside" the agent. Both approaches rely on the same two components: (1) a collection of video samples of the face from which to extract structured data about a subject's facial state and (2) a Neural Network (NN) controller trained from the aforementioned data, and programmed to control an avatar's facial mesh that resembles the face of the performing actor found in the collection of video samples.

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

A concern for accuracy is also shared by Affective Computing researchers. The process of validating the accuracy of facial emotion elicitation video corpora has primarily been used for research in the production of NNs for Facial Emotion Recognition (FER) software systems. The primary intention of corpus validation has been to warrantee that the emotion label value assignments for each frame for static images, or each clip for dynamic images, is statistically

consistent. The system of giving intensity values to emotion names identified in static photos evolved from a century-old method of recognition techniques.[133] Contemporary emotion recognition classifies facial muscle group behaviors into culturally and linguistically determined emotion names, or "labels".[134] The use of FERs provides a ground truth referent on which to model the facial expressions for NPCs. Developers of games and interactive media need a method to determine if the two components that influence the behaviors of an animated character or agent – the NN and the video corpus that trained it – are producing facial emotion elicitations as intended. Thus, video game developers of photo-realistic characters can draw from the techniques of both graphical and affective computation to determine the emotionally resemblant quality of their corpora and NN. Using some aspects of both approaches, a method of corpora production and evaluation can provide consistently evaluated data sources for training NN controllers. Two statistical techniques are proposed that provide a preliminary basis for analysis. used to train it.

5.2 Related Work

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

5.2.1 Example-Based Animation

New methods of simulating facial elicitation in Computer Graphics prioritize graphical accuracy of modelling and animation over emotional autonomy. Several studies by Paier et. al propose a "hybrid approach" that use "example-based" video clips for frame-by-frame facial geometry modelling, texture capture and mapping, and motion capture.[135] In their experiment, a performing actor speaks a few lines or elicits a set of idiosyncratically defined gestures. The recording or real time live stream provides information for automatic geometry and dynamic

facial texture generation.[136] Then, a NN using a variable auto-encoder (VAE) integrates motion for mesh deformation, while another NN selects animation sequences from an annotated database. Database annotation of animation has demonstrated the efficacy of movement data classification of a single actor that can be used later for semi-autonomous expressions utterances. Their approach demonstrates highly resemblant avatar animation for short single-word utterances or single-gesture elicitations.

An assumption that using speech as a primary modality for determining emotional states, belies the belief that facial elicitation is more reliably understood as a function of word utterance. The emphasis on speech synchronization assumes that the expressive meaning of an intended facial elicitation will correspond to the semantic context of the spoken word. This emphasis is found in a study by Suwajanakorn et al. that uses the vast collections of video samples of a U.S. President.[137] From a 17-hour corpus, the investigators mapped speech from persons who were not the subject of their video corpus, onto a moving and speaking face of the presidential subject. Their method discovered that optimal training of their NN benefited from expression positions of the face of both past and future video frames to best predict how to synthesize deformations of the mouth right before, during and at the completion of spoken utterances. Thus, their NN incorporated Long Short-Term Memory (LSTM) cells to predict mouth animation synthesis for the video of upcoming visemes. For the experiments conducted for this research, we also deployed LSTM cells and found them useful for the same benefit.

From the standpoint of a designer of autonomous agents for video games however, neither approach mentioned thus far provides a model of fully autonomous elicitation in response to measurable stimuli. Both examples show that the use of single-actor or single-subject corpora is viable for training a NN to simulate the facial expressions of an actor’s character design or that

of a real person. Viability is made possible with a NN that learns the dynamics of facial expression based on labelled visemes. This technique we also integrated through training with multiple video clips of an actor repeating a performance in reaction to the same stimuli. This approach proved useful in our development of NNs targeted to train specific emotions.

Another distinction in the Computer Graphics approach is that their corpora structures do not correlate with widely used psychological classifications of emotions and their elicitation (e.g. the six to twelve basic emotions identified by social and computational psychologists). Neither Paier's et al. nor Suwajanakorn's et al. research disclose a classification system of emotions, and therefore the meaningfulness of their synthesized expression relies on arbitrarily selected spoken semantics rather than independently systematized semantics of facial expression. The graphics approach instead prioritizes frame-by-frame facsimile of labelled visemes on a real human source as displayed by its simulated avatar. Nonetheless, the mimetic quality of results created by the workflows of both Paier et al. and Suwajanakorn et al. must be considered for autonomous facial emotion elicitation.

5.3 Production Methods

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

5.3.1 Corpus Production

There are two phases for our corpus production: first, designing a dyadic conversational scenario and second, rehearsing and recording video clips. Scenario design consisted of two characters for

actors to perform asynchronously following a *dialog behavior tree* in the form of a directed acyclic graph. Actors were cast and rehearsed in preparation of the video clip recording. One actor played the Stimulus Source character and recorded a video edited beforehand of all path variations as if they were addressing the other character, the Emotion Model. Then the Emotion Model performed back to the camera reacting to synchronized video from the pre-recorded performance paths of the Stimulus Source.

The design of the dialog behavior tree consisted of distinct fixed start and end nodes with three layers of six nodes in between. These intermediate layers allow two possible nodes of dialog turns. No node or edge could be repeated within a path of the tree as illustrated in Figure 5.1. The rules of the graph allow 32 paths through the tree. Each edge segment (the circle labeled letters) of the tree had a targeted emotion label. Segments used the same lexical content from the tree, though all paths had distinct dialog sequences. With the variations in paths, the dialog behavior tree created a permutable performance structure with stimuli for elicitations to occur.

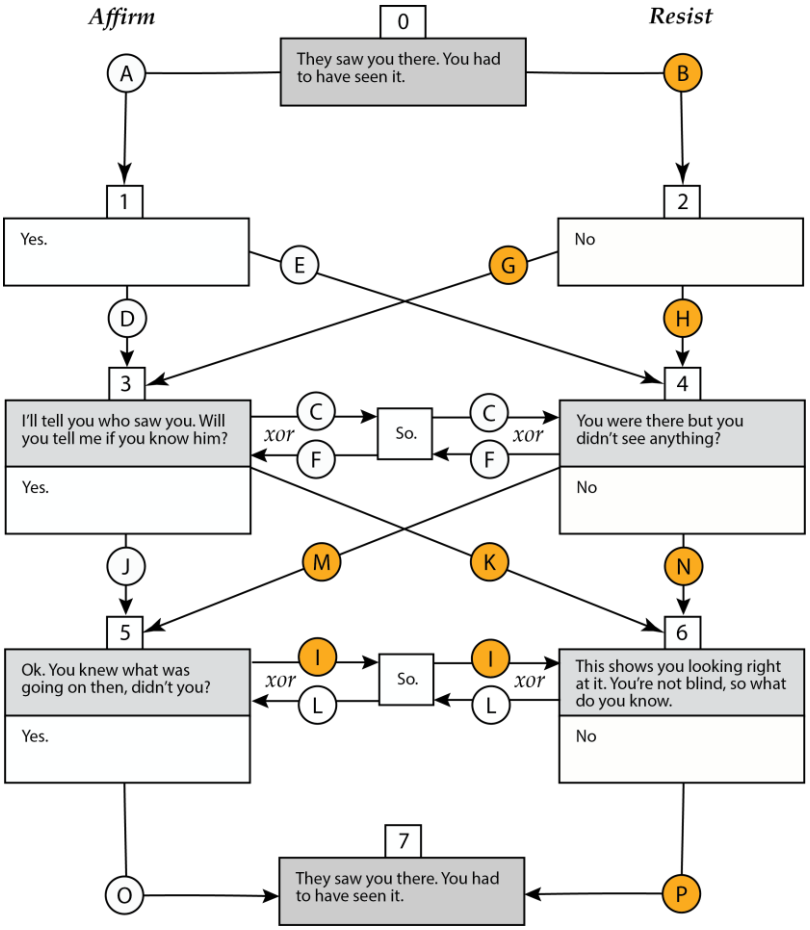

Figure 5.1: Box Nodes at Dialog Turns (3, 4, 5, 6) and Monolog Events (0, 1, 2, 7). Edges represent mental actions. Orange-colored nodes were used in the data analysis. As an acyclic graph, a path can use edge C xor F and I xor L.

Each path was recorded 9 times using three distinct degrees of intensity of action: high, medium, and low, to provide more variation to train the NN as practiced by Wingenbach et al.[138] Thus 9 clips times 32 paths yielded 288 total clips of the Emotion Model.

5.3.2 Post-Processing Emotion Analysis

Each clip was post-processed by the FER, *Noldus FaceReader 8*, for frame-based emotion analysis. The output data consisted of 7 normalized emotion label values for each of six emotions, *happiness*, *sadness*, *anger*, *fear*, *surprise*, *disgust* plus *neutral.* A new tuple of seven normalized values were output 3 times each second. Noldus FaceReader is a tested and ranked

FER system that has produced emotion recognition validation results that match the accuracy of human annotators.[139] Furthermore, the recognition accuracy rate of Noldus FaceReader has been documented as high at 94%. Its output is machine-readable text consisting of tuple instances of emotion recognition scores for each frame of video. These scores became the data used for training the NN using the Python language and the TensorFlow library. The FER was used to analyze all emotion values, plus neutral.

5.3.3 From Emotion Model to NPC Avatar

The Emotion Model was orthographically photographed from overlapping angles to produce a photorealistic head mesh that resembles the actor. The head mesh was generated by the *FaceBuilder* plugin for the 3d animation software system *Blender*. FaceBuilder is a modelling tool for supporting 3d head animation with a facial rig whose vertex groups are controlled by shape key actuators within Blender. These shape keys were designed to move the same alignments of facial muscle groups defined in the AUs of FACS. The head mesh and the shape keys embedded in the facial rig were deployed in the game engine Unity 2022, as depicted in Figure 5.2. The shape keys were put into autonomous motion by programmable blend shapes in Unity that receive streamed emotion data from the NN animation controller responding to the face of the Stimulus Source. The embedded NN receives the FER data and “reacts” to it in a way that intends to statistically resemble the character behavior the actor created in the video clips in the single-actor corpus. The NN-generates prediction data in the form of normalized emotion label values as a means of autonomously controlling the FaceBuilder head mesh to animate facial expressions.

Figure 5.2: Developing Emotion Model Avatar. Using Blender plugin Keen Tools FaceBuilder.

5.3.4 Recurrent Neural Network Architecture for NPC Facial Animation Controller

Among the clips generated for this research, 68.8% of the corpus (198 clips) was used only for training the NN. 20.1% of the corpus (58 clips) was used only for validation. The remaining 11.1% of the corpus (32 clips) was used to test the NN's behavioral resemblance to the actor corpora on which it was trained.

The principal components of the NN follow a Recurrent Neural Network (RNN) design. Each component of the neural network was selected for its probabilistic ability to choose values of coefficient weights and biases for specific input features of the data that the NN was trying to predict. Predicting the facial elicitation of game characters based on training data from an actor's performance requires spatial and temporal data representation. For our experiments, facial feature positions were estimated from their spatial contexts using Dense architectures (fully connected). We used a Dense layer of perceptrons that were fed two layers of bi-directional LSTM cells. The LSTM layers auto-regressively receive emotion data from 10 seconds in the past using 3 instances of data per second. Since the data for this experiment was fed pre-

processed emotion data tables (as opposed to a live video stream), the NN analyzed 10 seconds into the future as well. These temporal relations of elicitation events in the data were processed by LSTM cell layers, while spatial relations of facial features were handled by the Dense cell layer. Each emotion label was assigned its own NN, so the designed recurrent NN was cloned into a team of 7 NNs and trained on synchronized data generated from each elicited emotion from the Emotion Model and the Stimulus Source.

5.4 Evaluation Methods

The evaluation methods proposed provide the developer of NPCs with a quantitative process that measures behavioral difference. Optimally minimized difference in data can be interpreted as statistical *resemblance*. It is our intention to demonstrate statistically, that given the same or similar stimulus, prediction data from the NN can control animation that resembles observed FER-generated data of the human Emotion Model on which it was trained. The resemblance then depends on minimizing the amount of error between predicted behavior performed by the Avatar and observed data performed by the Emotion Model. But the predicted data is not a single set for each instance in time. Instead, each time-instance within a path through the dialog behavior tree is a video frame shared by at least 9 video clips and their edge segments, as well as other paths that share the same edge segment. Therefore, since all the video clips are precisely synchronized, each of the frames in the experiment has a mean emotion value drawn from at least 9 clips of the same edge segment. And this value can be used to calculate error in relation to the predicted value at that frame demonstrated by the NN. By looking at the difference within ±1 standard deviation, two useful statistical properties provide the results to determine statistical resemblance. By calculating the Percentage of Extreme Residuals (PER) that fall outside of ±1 standard deviation, a first test of resemblance can be applied to the prediction data. To support

those results, the Root Mean Square Error (RMSE) provides an amplification of variance. If the *neutral* emotion label values are used as a benchmark, RMSE becomes an additional statistical property to show if a NN facial animation controller resembles the character facial movement the actor generated, and if the corpus that trained the NN is sufficiently robust to confirm resemblance to any emotion labels.

5.4.1 Percent of Extreme Residuals

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

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

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

Validation of a NN facial animation controller should not require the facial mesh behavior to exactly animate the same way every time when it receives the same input from the

Stimulus Source. But how wide should a range of variance be to seem human-like? Consider ±1 standard deviation. With the mean of any emotion value at each frame as a reference, a range of *resemblance* can be defined around the mean by using the standard deviation. The experiments of this research found as resemblant the frames where the predicted emotion value of a given frame fall within ±1 standard deviation from the mean. For this condition to occur, the predicted value will fall into a value space with at least 68% of the observed values. For frames that fall outside ±1 standard deviation, they shall be determined as *not resemblant*. In essence, we interpret the percentile determination as a boundary indicating if the NN predicted behavior for a frame instance for more than 2/3rds of the clips in the corpus that traversed that instance in the dialog behavior tree which has its specific instance of stimuli from the Stimulus Source.

5.4.2 Root Mean Square Error

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

The Root Mean Square Error (RMSE), a frequently referred statistical measurement of difference between predicted behavior and observed data, is used to measure facial elicitation resemblance as a function of variance between two synchronized time series data sets: observed test data and generated prediction data. Each of the differences between the observed and predicted values, referred to as residuals, aggregates their magnitudes from point to point in a data set. The resulting value is always positive where 0 is a lower bound representing a perfect fit

between the observed and predicted data. The mean of observed values for each emotion at each frame was used as a baseline to compare amplified variance between the emotion labels. The RMSE is calculated as follows:

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

y_i is the observed value at the *i*-th frame. y_p is the predicted value for the *i*-th frame. *n* is the population count, or total number of values in the population of observed values. Under the radical sign, RMSE aggregates and amplifies the variance by squaring the summed difference before dividing by the population count *n*. A subset of squared differences will be magnified quadratically and summed once before being squared. Meanwhile, a subset of squared and summed smaller differences will diminish quadratically. Thus, the extreme differences, large or small, will be amplified. Since emotion values of video frames of facial emotion elicitations are normalized to fractions in a 0 to 1 scale, magnification of variance provides a visible contour of the behavior of residual values in relation to the mean of observed values. A low RMSE will lean toward greater *resemblance* provided the PER is also *resemblant.*

5.5 Results

Statistical error scores were computed to determine the behavioral resemblance between the mean of emotion values at each frame for all of 6 emotions. See Table 5.1.

TABLE 5.1: SUMS OF EMOTION VALUES

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

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

The highest summation of observed values should provide the lowest percentage of PER errors for the predicted values of emotion elicitations. As shown in Table 5.2, *neutral* PER is 0.0794, the lowest of all recognizable labels. Therefore, following the behavior of *neutral*, the next highest PER emotion may provide proportionally resemblant results, proportional in that the higher the percentage the emotion's sum is to the sum of neutral, the more resemblant the primary emotion values are in relation to the *neutral* scores. Table 5.1 shows the emotion with the highest percentage close to neutral for frame value sums is *sadness* with 81%. The next

highest is *anger* at 6.95%. With such a distant second position and so far off the benchmark of *neutral*, one should doubt the resemblance of *anger*, while taking note of *sadness*.

TABLE 5.2: ERROR FOR EDGES B, G, H, I, K, M, N, P

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

The next consideration is the *spread of emotion values* for each frame of each emotion. One may notice in Table 5.3 that *neutral* again behaves as a benchmark, evenly distributing values with a relatively smooth and centered distribution. *Sadness*, the emotion closest to neutral, while somewhat skewed to the lower half of the distribution pentile shows an even distribution. All other emotions are far less evenly distributed, with most of the values compressed into the first and second pentile of values.

TABLE 5.3: PROPORTION OF VALUES

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

Lastly, again consider the error data as seen in Table 5.2. Interpreting error requires the mean emotion values of any emotion to be high enough so that the region of the standard deviation is nearly all above zero. If the standard deviation region is clipped by a zero-value line, then the prediction values will likely rest above near-zero as well, providing no "bottom room" to dip below the standard deviation region. Unlike *sadness* as shown in Figure 5.3 and *neutral* seen in Figure 5.4, the predicted data for *scared, surprised, happiness* and *disgust* show unreliability for the NN and corpora for this research because their standard deviation regions drift over the zero-value line causing the PER and RMSE to appear to support accuracy, when in fact the NN is reacting with very little elicitation response for the given stimulus (flat lining). Most interestingly, *sadness* shown in Figure 5.4 and to some degree, *anger* depicted in Figure 5.5, show some promising responsiveness to the stimulus, reacting in similar ways as the mean of the frame of emotion values in the observed test data as indicated by the PER score for both in Table 5.2. With the RMSE score for *sadness* at 0.2438, its score is 0.0581 higher than neutral at .01857. *Anger* shows a lower RMSE, but anger values are still too low to be fully reliable with much of its standard deviation clipped by the zero line and the observed values of the test data also dropping to zero for nearly 20 frames.

The difference between the RMSE and the meanRMSE is that while the RMSE score looks only at the difference between the mean of the observed test data and the predicted data, the meanRMSE is the mean of all the plotted RMSE scores shown in red in each of Figures 5.3, 5.4 and 5.5. The plotted values show the RMSE for the chosen emotion at each of the 126 frames examined in relation to the same synchronized frame in the predicted value. The fact that the two RMSE scores are close in value provides a check on the accuracy of the error assessment process.

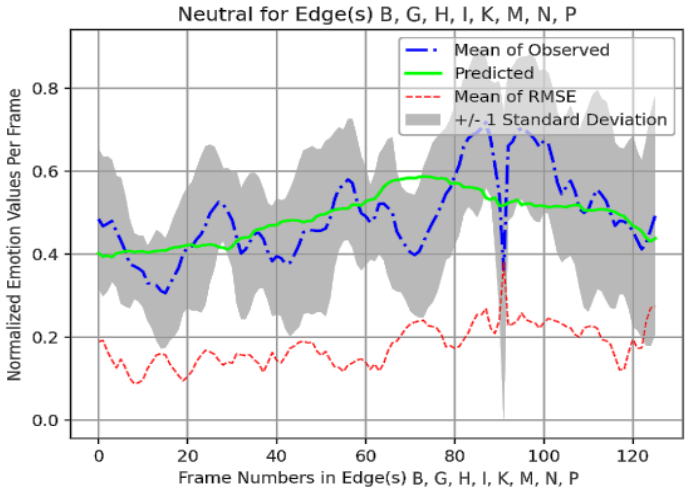

Figure 5.3: Neutral Benchmark, Observed Against Predicted Values. Used to compare with other emotions.

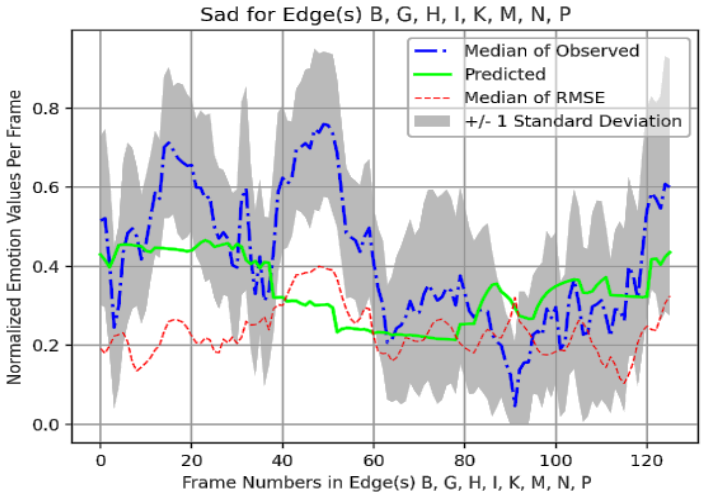

Figure 5.4: Sadness, Observed Against Predicted Values.

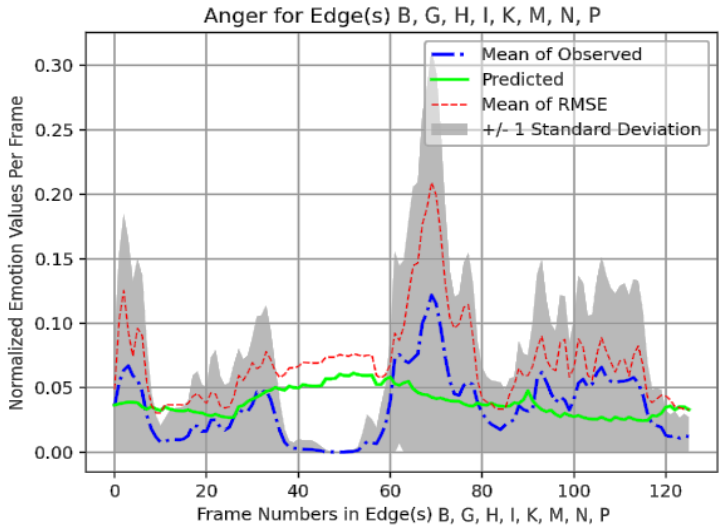

Figure 5.5: Anger, Observed Against Predicted Values. Demonstrates marginal resemblance with possibly too little responsiveness to be reliable.

One inexplicable anomaly is an apparent minimum value PER of 0.1984 in the test results shown in Table 5.2. Mathematically, it has been determined that for the 126 frame results for each emotion, 25 frames fell outside the standard deviation for the three emotions: sadness, fear and surprise. It remains unclear if this fact is a coincidence or caused by the test design.

5.6 Conclusion

Thus far, this chapter has identified several statistical properties relative to neutral: RMSE, PAR, and Spread of Emotion Values. What has been demonstrated is that at least one emotion label, sadness, was successfully simulated. What might be useful is a classification system for each of these measurements that would provide discrete labels within a range. Such a classification system could indicate if the results will lead to a NN model that will output predicted emotion values that range from resemblant to not resemblant in relation to those values elicited by its human actor referent. The aim of this research is to expand the creative process of character design for video games beyond the modeler and animator and toward the skills of the actor. For

the methods proposed to become useful, they must also produce salient results that confirm resemblance. Thus far, this research has demonstrated a statistical method to validate resemblance. Further investigation should confirm its viability as a method of production for game character production workflow.

6 MEASURING EMOTION VELOCITY FOR EVALUATING RESEMBLANCE[5]

NN animation controllers can enable actors to collaborate with game designers during the production of a video corpus of a character's behavior in a game scenario. This contribution to the authorial process of animated character behavior needs measurable validation techniques to determine if a corpus' sample collection or NN architecture 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 analyze 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.

6.1 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,[140] 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 has demonstrated many breakthroughs with the use of neural networks trained from live action video subjects to

[5] Schiffer, S. (2023). Measuring Emotion Velocity for Resemblance in Neural Network Facial Animation Controllers and Their Emotion Corpora. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence - Volume 1, SciTePress, pages 240-248. DOI: 10.5220/0011676200003393

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.[141] Therefore, quantifying and comparing velocity can 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 NPC's 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[142] and seek to classify emotions through mostly static facial expressions as systematically practiced by psychologists Ekman and Friesen[143] and continued by Ekman.[144] 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.[145]

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 requires 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 chapter 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 a FER of recorded facial expressions. We compare the velocity of data generated by a 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.

6.2 Related Work

Past research concerning autonomous facial animation comes from the computer graphics and affective computing communities. Each contribution prioritizes distinct objectives frequently espoused by their respective research communities. Computer graphics evaluates experimental results with the expectation of high visual resemblance and positional accuracy of animation. Affective computing evaluates a synthetic agent's behavior in relation to a psychological model of behavior and its effect on human users.

6.2.1 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.[146] 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.[147] 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.[148] and a study by Suwajanakorn et al. that uses the vast collections of video samples of a U.S. president.[149] 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.

6.2.2 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.[150] Further research suggests that perceptions of "naturalness" were greatly affected by changes in expression velocity.[151] 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.[152] 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.

6.2.3 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.[153] and Benda et al.,[154] versus non-actors.[155] Similarly opposed approaches is the choice to use tightly scripted scenarios as did Busso et al.,[156] versus more improvisational techniques.[157] Lucey et al. proposed to record video in a lab-controlled environment,[158] while others collected samples from major media production industries such as news and entertainment.[159] 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.[160] 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.[161] 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.[162] 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.[163]

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.[164] 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.

6.2.4 Neural Network Architecture 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.[165] The Method approach, a series of practiced exercises as developed by Russian theater director Constantin Stanislavsky,[166] were combined in the Oz Project with the emotion system structure of Ortony, Clore and Collins, also known as Appraisal Theory.[167] 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 intentionally 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.[168] Loyall et. al produced several interactive media experiments with autonomously animated characters using emotion AI systems that deployed digital models of emotion.

Implementation architectures have continued to progress for autonomous 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.[169] 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.[170] In its earlier versions in social and educational games, the FAtiMA architecture had proven to be effective for learning and engagement.[171] 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.[172] Unlike the method that this research proposes, these previous related works did not use single-actor video corpora.

6.3 Methods

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-behavior tree authored in the form of a directional acyclic graph as seen in Figure 6.1. 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.

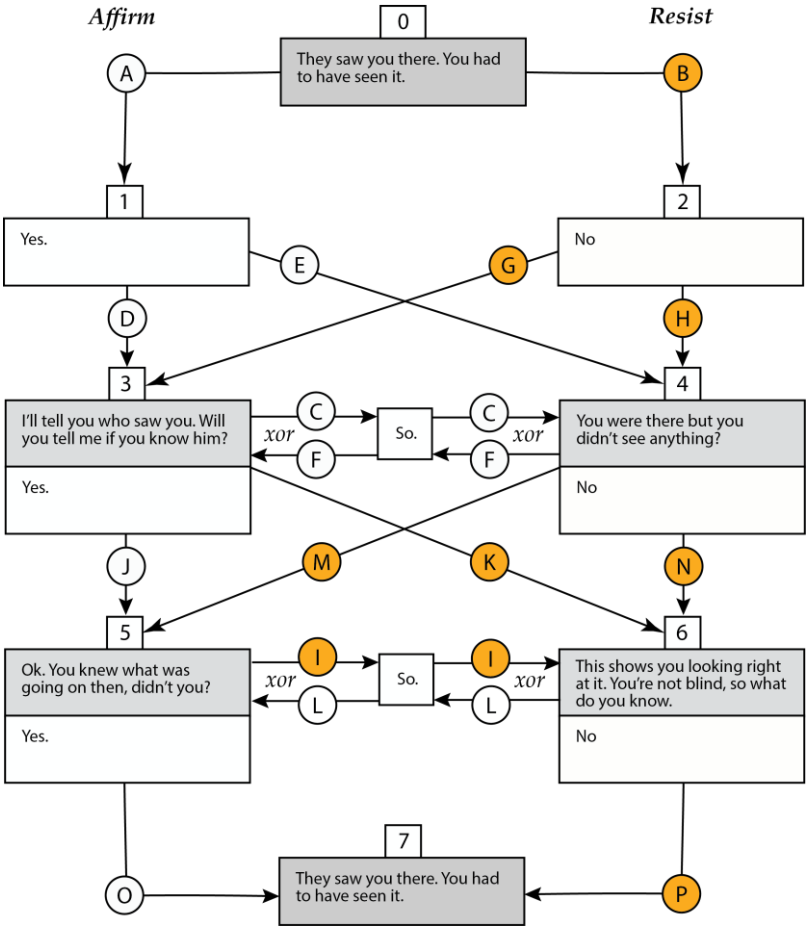

Figure 6.1: Dialog Behavior Tree as Acyclic Nodal Graph. Circled labels are edge segments. Orange-colored were used in the experiment. Path can use edge segment C *xor* F and I *xor* L.

6.3.1 Producing the Corpus 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 6.1, 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 6.2.

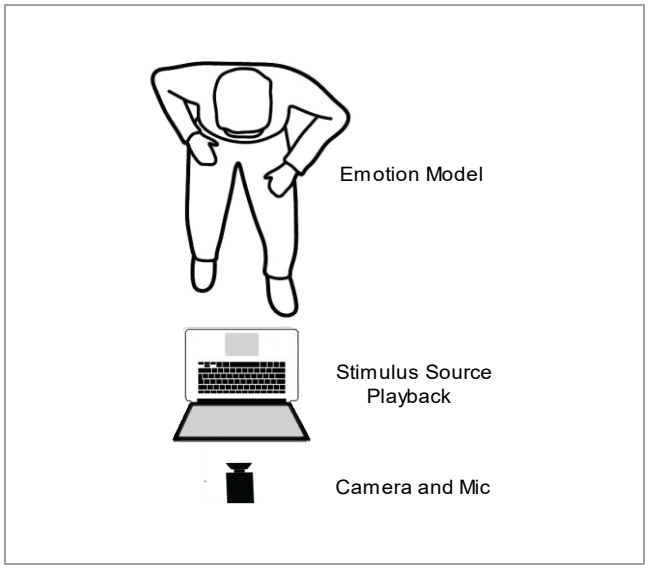

Figure 6.2: Production Setup. Emotion Model sample recording for facial emotion expression corpus.

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.

6.3.2 Modeling 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 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 actuators 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. A frame sequence of both the Stimulus Source and the Avatar in Figure 6.3.

The NN-generates its predictive data as the reaction in the form of normalized emotion label values as a means of autonomously 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.

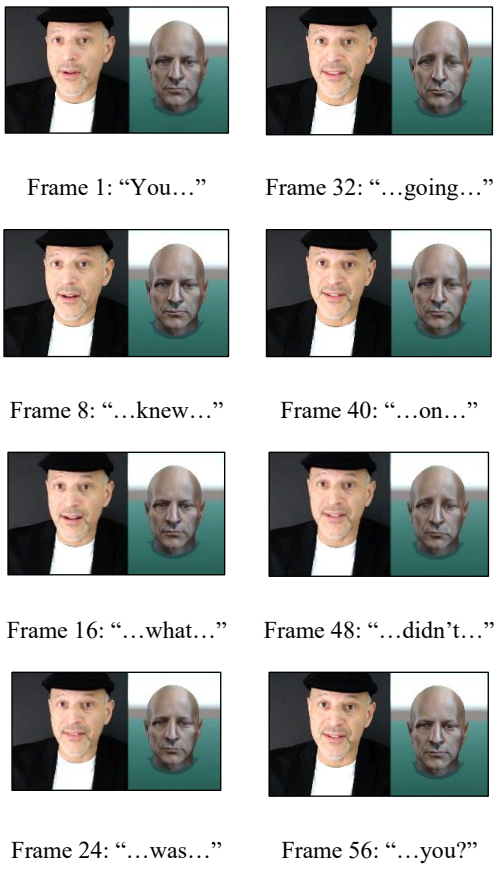

Figure 6.3: Eight Synchronized Frames Over 2.3 Seconds. Shows Stimulus Source (left) providing action: *accuse*. NN Avatar (right) reacts with sadness.

6.3.3 Modeling 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) were 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.

6.3.4 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 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.[173] Furthermore, the recognition accuracy rate of FaceReader has been documented as high at 94%.[174]

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 performed 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.

6.3.5 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 chapter, 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 6.1 (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 were 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 consists 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:

$$e_{\mu} = \frac{\sum_{s=1}^{S} e_{\mathrm{k}}}{|\, S \,|}$$

The quotient e_{μ} is found by finding the sum of values e at frame k for every clip that traverses frame k for each emotion, and dividing by the count of frames in set S. 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 6.1. 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}.

$$v_{\mu} = \frac{e_{\mu t41} - e_{\mu t0}}{t_{41} - t_0}$$

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 controlled 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 below.

6.4 Results

Figures 6.4, 6.5, 6.6 and 6.7 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 as seen in Figure 6.4. 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 6.1 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.

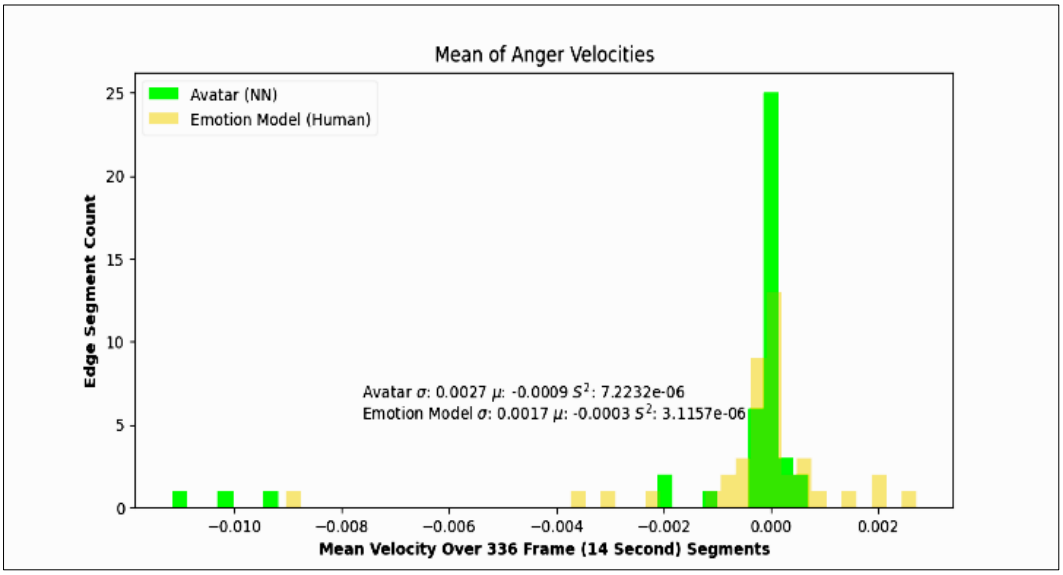

Figure 6.4: Distribution of Anger Velocities.

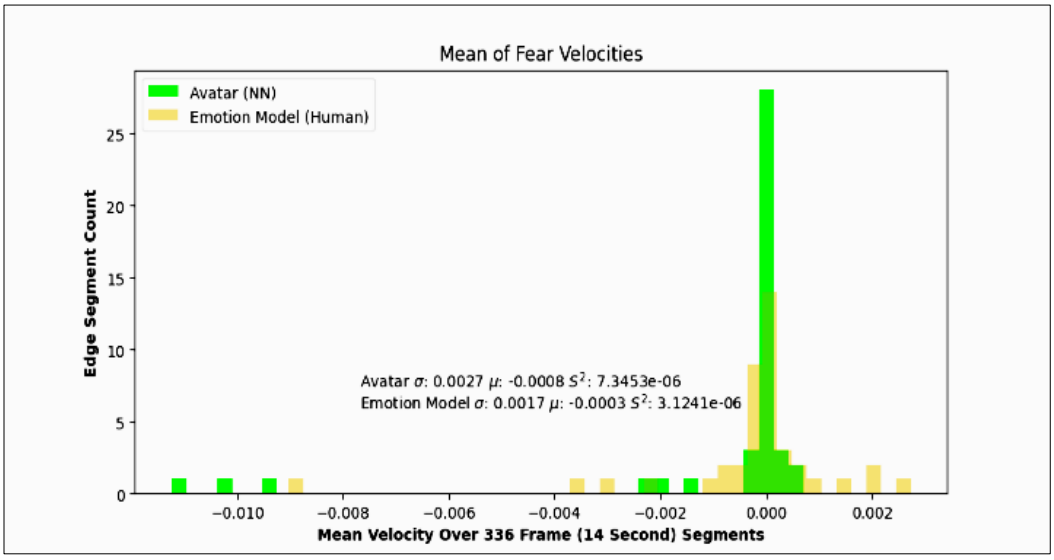

Figure 6.5: Distribution of Fear Velocities.

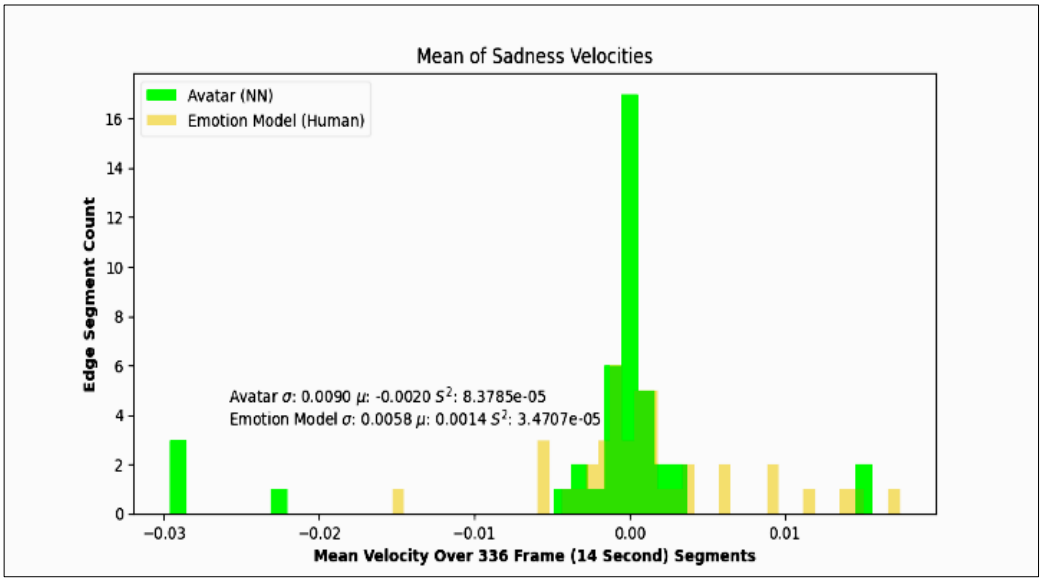

Figure 6.6: Distribution of Sadness Velocities.

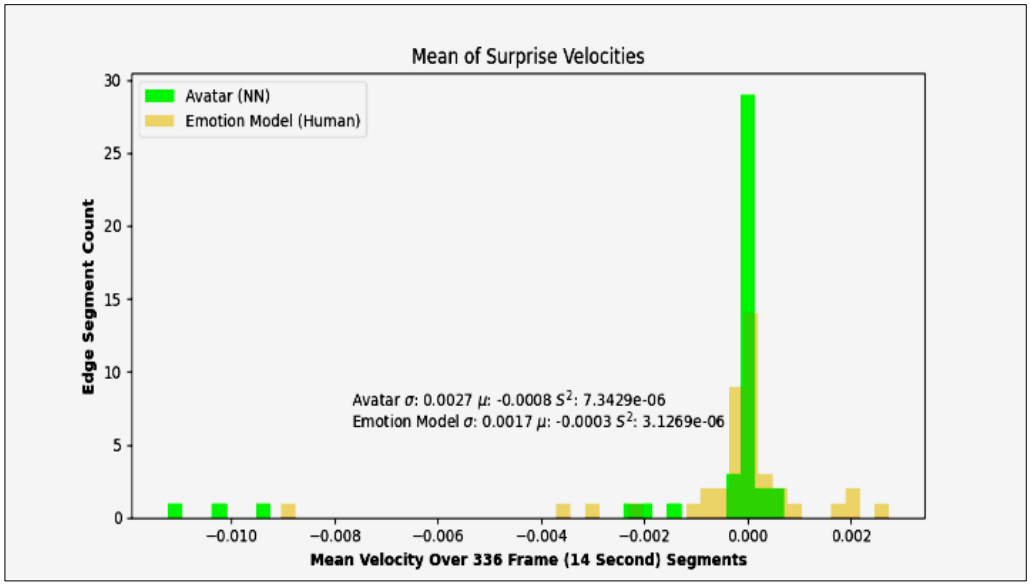

Figure 6.7: Distribution of Surprise Velocities

TABLE 6.1 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

6.5 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.

7 ANNOTATING FACIAL EMOTION CORPORA FOR VIDEO GAME NON-PLAYER CHARACTERS[6]

This chapter proposes actor-centric methods for corpora annotation using actor performance theory as part of a basis to interpret the labeled data derived from facial emotion recognition software analysis of a single-actor corpus of facial emotion expressions. An actor-centric method empowers the performer to become a creative partner in the design of their own representation in a video corpus and defies the trend in AI applications of positioning artists as pattern-making subjects ripe for exploitation by AI practitioners looking for unauthorized content to feed their neural networks. Emotion data captured during corpus production is used for training neural network animation controllers that animate the face of humanoid characters for photo-realistic rendering of cinematic facial animation.

7.1 Introduction

The reaction among some in the arts communities to autonomous machine-derived creative production has ranged from skepticism toward its aesthetic value, to hostility toward its ambiguous and potentially intrusive role in the workflow of artistic production. Much of the skepticism and hostility has been a response to implementations of AI that segregate artists away from the design of the sampling process for machine-learning, and instead positions artists as only producers of content whose patterns of expression can be freely sampled by machine-learning algorithms with neither the artist's consent nor participation in the sample creation process. Granted, much of the artistic product used in the most current iterations of AI-derived artistic product has been from artists who are long deceased. But such a limiting and potentially

[6] This chapter is under consideration with Frontiers in Computer Science as a part of their series on Human-Media Interaction under the research topic Corpora and Tools for Social Skills Annotation.

exploitative relationship could only be a formula for resentment and decline of innovation. At the time of this writing, AI-derived creative products generated in the style of graphic or literary artists has been the primary subject of global fascination and acrimony. It is only a matter of time before the performing arts of music and acting become new material for autonomous machine emulation. Eventually, they will sample existent digital corpora of deceased or living performing artists to train neural networks (NNs) to play virtual musical instruments or to 3d modeled characters to appear and play as known actors or characters. The current model may exploit large samples of music or acting and eventually yield performances that satisfy audiences. It may also disturb audiences with the awareness of the machine-derived source of the artifice. The passive role that positions artists as an exploitable resource for technology is neither inevitable nor ethical. Artists can master the process of designing corpora as they have many other complex media, thus making the audio or video corpus itself a new medium of artistic expression intended for machine-learning algorithms that collaborate with an artist's intentions.

A starting strategy for this approach is found among the art forms that already have a well-developed theoretical foundation that organizes the training and preparation of the performer toward a tangible performance, while using iterative techniques of structured repetition. This research examines the annotation technique often used in the preparation an actor makes for a role in theater, film, and television for the purpose of creating a corpus of video clips that will be used to train a NN that will animate a photorealistic 3d non-player character (NPC) model. The NN ultimately controls the animated facial expressions of an NPC performed by the actor in-character. Annotation as preparation for theatrical or screen performance often demonstrates how an actor uses their internal emotional behavior as an impetus to create

character-specific movements, postures, gestures, and facial expressions – all embodied elicitations.

In addition to the performing arts disciplines, the field of psychology has also investigated and theorized extensively the human process of emotion generation. Fictional performance of facial emotion expressions communicates (albeit somewhat ambiguously) a character's "appraisal" of three perceivable categories of objects in terms that either affirm or interfere with a character's beliefs, desires, and intentions. Those objects are *events*, *agents*, and *objects*. These three components are the principal stimuli for emotion generation as defined by Appraisal Theory, a widely referred psychological model used to describe the feedback system that generates emotions within humans. Appraisal Theory explains how emotions are generated in human interactions, in terms that describe emotion as observed phenomenon occurring in the course of "real life", as opposed to the artifice of life as experienced in theater, film and television.

Performance Theory and many of the practices developed by or derived from the Method of Stanislavsky and his many disciples, focus primarily on the artificial production of emotion used to train actors and prepare them to portray fictional roles for performance. As we will discuss later, Appraisal Theory and Performance Theory refer to the same human system of emotion generation, but for distinct purposes. The centuries-old Performance Theory developed as a means to train and guide actors creating fictional characters. Appraisal Theory evolved to both describe a system of human emotion generation, and to provide a blueprint for simulating human emotion in a computer. Inevitably, the two theories can be considered together for the creation of autonomously emotive humanoid characters, either in the form of a 3d model for a computer-driven interactive experience, such as a video game or conversational agent, or for a

life-sized android machine in humanoid form. Psychology vis-à-vis Appraisal Theory provides a model for generating emotional *responsiveness* in synthetic agents – generating physical elicitations that appear to externalize the feelings we experience when something happens to humans. Theater and screen media vis-à-vis Performance Theory provide a model for emotional *agency* in artificial agents – generating the feelings humans experience when engaged in an action to get a goal. Responsiveness (from external entities) and agency (toward external entities) are thus two fundamental poles for elicitation that determine which template to refer to when annotating a script for production (planning agency), or when annotating a video clip within a corpus (documenting responsiveness).

Annotation as a process of planning or documentation has great value in that it organizes the thoughts of the performer and the perceptions of an observer of the performance around principles that motivate, classify, and quantify an actor's behavior as their fictional character. Annotation as preparation for performance aids an actor at filtering and internalizing their real time stimulus so that when performed is live, it can be appraised automatically in character, and not as themselves. Once internalized, an actor can then take one of a set of possible actions that belong to a character's action set. Annotation aids the observer so that they can describe the emotional quality of a performance using commonly understood natural language emotion labels (e.g., happiness, anger), visemes of elicitation (e.g., smile, scowl), and percentile proportions that describe how close a performer is from maximizing an elicitation of a particular emotion label (100%), or how close a performer is to showing no elicitation of any emotion label at all (0%).

As technology progressed toward simulating human forms and movements using 3d polygonal meshes of varying resolutions, simulating human facial expression with computationally enabled hand-crafted animation became a standard artistic practice through the

late 20th and early 21st centuries. Referencing emotion annotation methods for 2d, and later 3d animation, was a common practice that relied on late 19th century theories of elicitation developed for oratory and theatrical performance. The rise of AI technologies that sample large datasets of human expression and the inevitable integration of NN driven animation of 3d facial meshes, will require a distinct technique.

The purpose of this research is to introduce a method for future animation development of NPCs in video games. This research may also inform the computational social scientist who may design and study autonomously animated performance. The result will be a model for annotation that integrates methods used in the preparation of an actor's character and role in animation – the *agency* of emotion – with the methods used in the modeling of human emotion in the natural setting – the *responsiveness* of emotion.

7.2 Related Work

First, we examine the original ideas set forth by performing artists and social scientists whose primary interest was to document, theorize and advance their specific disciplinary interests of either cultivating efficient creative production of actor performances for stage and screen media, or to more deeply understand how the human emotion system works. These artists, theorists and practitioners are selectively summarized. We also examine related work of researchers and experimental artists who used the relationship between the performing arts, psychology, and computer science to model and design autonomous emotion generation systems. Lastly, we consider the technology of emotion recognition that autonomous emotion elicitation systems rely on for machine learning. The research surveyed has made significant contributions to our work through overlapping fields that needed annotation to classify, quantify, and simulate emotions.

7.2.1 Emotion Generation and Annotation Methods from Performance Theory

One late 19th century stream of discourse partitions theater with oratory and embraces the idea that to persuade an audience of the believability of a play, much like a speech, an actor must be self-aware of their own body's expressive meaning through its emulation of a taxonomy of expressive gestures and postures. The presumption was that for emotional ideas to communicate, the position of the body must be legible. French orator and performer Delsarte formalized a system of expressive gestures that intended to evoke emotions in the audience and correlated them with corresponding emotion labels and the states of being they inferred.[175] The legacy of the Delsartean premise was firmly embedded in the careers of silent screen American actors, such as Lilian Gish, who transitioned their acting training for stage into a working method for the screen.[176]

In the early 20th century, the identification of emotions as a lexicon of labels with corresponding body positions is found also in Russian theater director Meyerhold. His Biomechanics system took on a similar premise, that emotions in the body use its appendages as actuators to elicit and communicate in a modality independent of spoken words. Both Delsarte and Meyerhold considered their work innovations of science and art whose structuring and classifying techniques would inform both the performer in their process and the study of emotion in the larger laboratory of life.[177,178]

Attempts to computationally implement and test aspects of the Delsartean expression system for use with conversational agents have shown promising results, specifically the use of space around the head with pose and hand gestures.[179,180] Computer animation research has also implemented techniques drawn from theories of Meyerhold concerning balance and the emotional meaning of its absence.[181] While the results from Meyerhold are not necessarily

"realistic", the process provides a means to classify emotion expression in the body using a system of physical practice.[182]

The techniques mentioned thus far prioritize the perception of emotion over its actual lived experience. Neither Delsarte nor Meyerhold and their disciples objected to an actor experiencing the emotion they intended to communicate, but their mission was to clarify for actors that the audience seeks recognizable visual forms in the performer's body. Whether a performer felt the emotions they communicated or not was less important.

As professional theater training evolved in the late 19th and early 20th centuries, the nascent science of psychology framed questions that overlapped with actor training. Psychological theories and limited experimental research were focused on psychoanalysis for the English, French and German speakers. Early psychoanalytic theory saw emotion elicitations as clues that could reveal more about the unconscious drives and desires then about a system of responsiveness to the day-to-day reactions to life's joys and disappointments.[183] While Freud laid out a structure for a theory of mind, we do not know that Stanislavski in Russia ever read it. Evidence indicates he heard (though may not have read) about Ribot, a psychoanalyst who wrote about the memory of emotions, or what Stanislavski and Ribot called "emotional memory", and then later quoting Ribot as calling it "affective memory."[184] The idea that an emotional memory was a mental object in all humans gives the opportunity for an actor or acting teacher to talk about them, classify them by label and intensity, and associate them with autobiographical narratives affiliated with a collection of remembered personal sensations.

As his teachings and practice advanced, Stanislavski realized along with some of his students, that "emotional recall" of memories alone was time consuming and produced inconsistent results that took the actor out of the sensations of a real time stage play and into an

affective bubble, less responsive to the events on stage. He developed a revived focus on the circumstantial conditions of characters that would require a systematic "method of physical actions"[185] to bring out of the actor the appearance of authentic emotion and allow a "conscious means [to] reach the unconscious"[186]. Through experiments with physical action rehearsal, the actor could develop knowledge of their character's (1) relationship with others and their environment, (2) material and emotional super objectives within the story world, within a whole scene, and within thrusts of action smaller than a scene, (4) realization of obstacles that can interfere or aids that can assist, and finally, (5) sensations of imagined outcomes in the event of failure and success toward a super objective. The teachers and theorist who continued Stanislavski's work in the United States prioritized that the actor prepare the forementioned to create a unique set of actions (verbs) that the character could use to possess their super objective. Those actions used in a performance should consist of a "through-line" that leads to finally obtaining a super objective or ultimately failing.

Among about a dozen renown disciples of Stanislavski who have structured his notes and theories into teachable methods, we find a frequently referred well-structured list of questions by Hagen that organizes the "data" of the Method-originated techniques to prepare for a role: (1) Who am I? (2) What are my circumstances and possible consequences, (3) What are my relationships, (4) What do I want? (5) What is (are) my obstacles, (6) What do I do to get what I want?[187]

While for an actor there may not seem to be an obvious need to clarify which the parts of speech, verb tense or point of view should be used to answer these questions in writing or speech, for a computational solution, grammatical specificity allows for the development of consistent design patterns that can be implemented in a synthetic agent. As performance theorists

and practitioners did not foresee computational applications of their theories, the task of structuring and classifying the fictional data for a character design was left for game developers to sort many decades later. Later, we will elaborate our application of Performance Theory into an annotation system that feeds into a data structure for an NPC.

7.2.2 Annotation Derived from Appraisal Theory

While in the early 20th century the classification of emotional memories was within reach of the Method acting theorists, that project was not pursued. This task was left for psychologists to pick up five decades later with Appraisal Theory. Psychologist and emotion researcher Arnold first queried how do humans differentiate emotions.[188] Her answers were evolved into a theory from the experimental psychologist Lazarus, who formulated a theory of emotion differentiation and regulation as a process of coping with stress.[189] These initial investigations among psychologists sought to describe a cognitive structure of emotions. The pursuit later evolved into addressing the dynamics of emotion, the neurobiology of emotion, and the interplay among individuals within their environment[190]. To accomplish these goals, psychologists developed annotation with a hierarchy of affect labels, a map to position semantic constructions of affect on a Cartesian plane, and a flow graph that describes the processing of emotions as a cycle.

As the mid-20th century progressed, theories of emotion were more numerous and quickly modified by their proponents. The structures and language used to describe them were not objectified by contemporary computer scientists until two psychologists, Clore and Collins, collaborated with computer scientist Ortony.[191] Their research provided a computable model of emotions that was intentionally designed for implementation in synthetic agents. Its schema has also been used to model the emotion generation processes of cultural systems, such as politics

and media. Collectively referred to as OCC, the three authors' ideas are pertinent to the design of synthetic agents. Their focus was on the valanced cognitive representation of things a person perceives. If any perceived thing can be evaluated on a continuum as *good* (positive) or *bad* (negative) with respect to that thing obstructing or assisting one's goals, or affirming or offending one's tastes or beliefs, then the signal of that perception is quantifiable and therefore computable on a one-dimensional axis called *valence*. OCC's conception of perceivable things was divided into three categories of construals: *events*, *agents,* and *objects*. These categories are commonly referred to in other disciplines concerned with theories of mind, such as philosophy and literature. These categories allowed for representation in object-oriented languages, which in the late 1980s, was just beginning to become a model for design patterns of the software development industry. *Valance* is a measurement of hindrance or help that a perceived thing provides toward achieving a desired *goal,* meeting a standard of *beliefs* or fulfilling one's individual *tastes*. Valence measurements imply a potential change in proximity to physical *goals*, and thus is quantifiable. Adherence to *beliefs* or satisfying *tastes* are each computable with logic if the language representation of *beliefs* and *tastes* are semantically logical and without contradictions of inclusion. Finally, the OCC model describes the emotion system as a cyclical process where stimuli enter the system, are appraised through filters of *goal assistance*, *standards evaluation,* and *taste satisfaction*. Then the appraisal is acted upon, and a new set of stimuli can enter the system.

When the signal of perception is converted into an emotion and classified as an *event, agent,* or *object,* it is "appraised". In the process of appraisal, some initial "undifferentiated" emotions are very transient, like *pleasure* and *displeasure*, *approval* and *disapproval*, *hope* and *fear* or *liking* and *disliking*. More lasting and well-differentiated emotions fall into six

categories: *fortunes of others* (resulting from events), *prospect-based* (appraisal of a future outcomes), *well-being* (of oneself), *attribution* (of an agent's action and its outcome), *well-being and attribution compounded*, and *attraction* (toward objects). We find a reprint of the OCC model of appraisal in Figure 7.1.

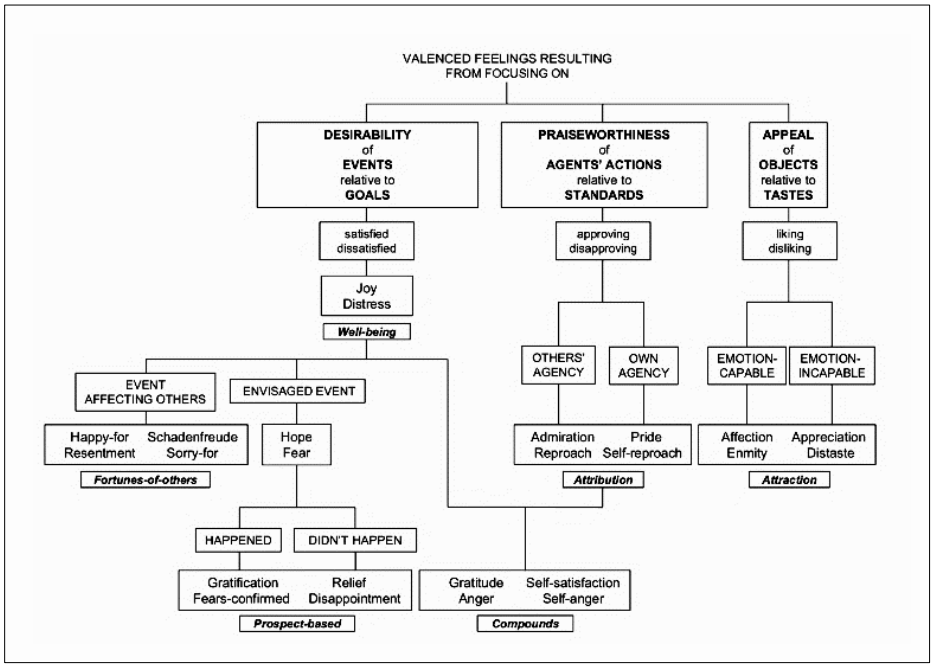

Figure 7.1: The Ortony, Clore and Collins Emotion Model of Appraisal. Describes three categories of construals and emotion types.[192]

We can trace the cyclical aspect of the OCC model from the point of view of an NPC. Stimulus comes from the game level at intermittent frames of time. The perception is recognized as one of the three construals. If the properties of the event, agent or object fit the sought for appraisal variables of the NPC within range of the required valence to trigger an emotion response, then a signal traverses the paths of the graph. Some paths are more driven by reason, and others by impulse. The model does not show how an emotion label is transformed into an elicitation. (This topic we discuss later.) But the resulting elicitation becomes an event itself in the environment for other agents to observe and react to, and in some cases, objects may become

altered (such as chairs thrown, or drinks spilled) due to emotion-derived actions. The cycle repeats when the elicitation or an effected agent or object creates new perceptions for any agent in the space to perceive (including the self).

One might ask, how do events, agents and objects get their valence values in a video game? The declaration and assignment of appraisal variables must be determined by the idiosyncratic design of the video game story world, which includes the circumstances of the NPC and their causally and narratively determined goals, predisposed beliefs or tastes. If we consider the discussion of Hagen above, we adjust appraisal variables dynamically as the circumstances change, as the proximity to goals change, and as appraisals of surrounding things adhere to predisposed beliefs or preferred tastes. For example, an NPC with a *vitality* property will modify the appraisal values of perceived events, agents and objects that relate to food and rest if they are hungry and tired. Any perceivable events, agents or objects can be recognized, evaluated, and quantified for their valence properties if an NPC is programmed to sense them.

The OCC model has three levels, of which two infer stages of reappraisal for each observation before a more lasting emotion label is processed and affect is derived. The first reappraisal occurs when an NPC decides if an event has consequences for itself only or for others only (*fortunes of others* versus *well-being*), or a combination of the two. There is also a path where the NPC assesses that an event is caused (*attribution* of agency) by others or themself, and then judges that agent's behavior to trigger an emotion elicitation. And finally, there is a stage where one reacts to an event by reappraising its effect on the future (*prospect-based*). The simplest of paths through the OCC model is the reaction to objects (*attraction*). Object appraisal is somewhat isolated from the other stages. OCC remind the developer that objects can become agents if the agent NPC perceives that an object has agency, even though objects actually do not.

This is the case of the object that seems to have a life of its own, especially one that obstructs an NPC's goals.

The introduction of the OCC model in the late 1980s provided limited opportunities for immediate implementation, as robotics, game development and other agent system designs were still rudimentary. In the1990s, the Oz Project, a collection of video game experiments and research papers realized by Loyall, Bates, and Reilly, made use of both Performance and Appraisal Theories to implement virtual agents in a digital artifact.[193] Some of the theories of the Method were implemented in the Oz Project with the emotion system structure of OCC.[194] The Oz Project's architecture for an agent called *Em*, consisted of several interactive media experiments with autonomously animated digital characters using emotion AI systems that appraise and elicit using much of the OCC model.[195] While the Oz Project made use of a theoretical foundation for motivating emotion elicitation, it did not use a dataset from which to draw machine learnable samples into a neural network model.

As microprocessors became smaller and faster, the possibilities for emotion AI also expanded. Cameras and sensors became more refined, interactive multimedia programming tools, like game engines, became more available, and data processing languages and computational methods similarly advanced. Emotion model implementation in the 21st century became more feasible. A revision of the OCC was offered by Steunebrink et al. to clarify and modify the OCC model so that its design would adhere to object-oriented design patterns for a software implementation.[196] The Steunebrink et al. revision has a similar flow in the emotion production process model as the original. The object-oriented revision makes computational implementation more obvious as seen in Figure 7.2.

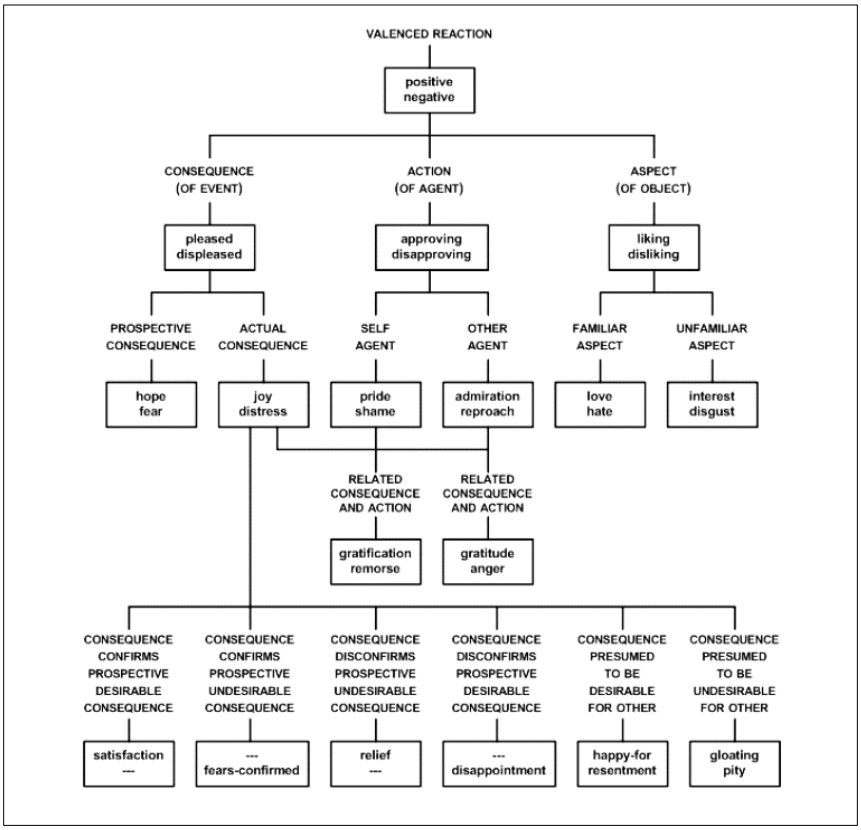

Figure 7.2: The Steunebrink et al. Revision of the OCC Model. Modified for object-orientated software design patterns.[197]

Object-oriented design allows for an “inheritance-based hierarchy”,[198] but it also imposes a rigor on the parts of speech used to describe the nodes and edges of their tree graph. The first row of the Steunebrink et al. model consists of predicate labels. Predicates are descriptive states, but they are not noun concepts. Predicates describe a transient state-of-being after a short judgment, but they are not emotion objects that can be possessed and encapsulated. Predicate objects allow for undifferentiated appraisal to be further distinguished after reappraisal. *Pleasure*, *displeasure*, *approval*, *disapproval*, *liking*, or *disliking* are thus cursory. And for Steunebrink et al., they are intended to be super classes that can be used for inheritance by subsequent derived classes that contain more detail and inheritance along an appraisal path. Each

of the parts of speech that are nodes of emotional states in the second, third and fourth rows are noun-object emotion concepts that a mind can possess. *Love, hate, gratitude, anger,* and *relief* are all possessable noun-object emotion-concepts. The exception is *gloating*. One might argue that it is a gerund, and therefore a noun. But here is an example of a language's limitations to create words that fully objectify a felt concept. The verb *gloat* expresses antipathy, but antipathy lacks the selfish glee of feeling pleased with another person's failure, as gloating implies. So, gloating remains an exception. Following the basic rule that emotions are objects, nouns make possessable emotion labels. Noun-object emotion concepts clarify for actors the sense of *choosing an action* to *possess the emotional object*. For example, *I beg to get relief.* It is the act of trying to possess the emotional object that is obstructed by an obstacle that gives clarity to the observer that an emotion is present. These are emotions resulting from NPC agency that we described earlier, that are distinct from emotions derived from responsiveness, and that have been called "the reaction in counter-action" in Performance Theory.[199]

Differentiation between emotions is crucial for validation tests to confirm accuracy. Too many or too few emotions can create problems for annotators to recognize and differentiate each one. Debate continues over the number of emotions. Theorists also disagree on which ones should be categorized as "basic", or if a basic set exists. The value of defining a short list of "basic" emotions is that a finite set is more comprehendible and commercially viable, regardless that such a list is not agreed. There is also much disagreement over whether emotion elicitations are socially learned or biologically universal among humans and primates. These issues present high stakes for software developers and the industries that have begun to depend on emotion recognition software.

Annotation must also be considered for the design of the internal processor of an emotion model. We considered for our annotation method the FAtiMA model developed by Dias et al.[200] Most important is its description of the Appraisal Frame. It is the intermittent fraction of time that the NPC appraises incoming stimuli and assesses the values assigned to properties of events, agents, or objects that it construes. In Figure 7.3 we find how appraisal variables are assigned numeric value types from an Appraisal Component. If the value falls within a range that triggers affect, an Affect Derivation results, which for a facial animation controller is synonymous with an elicitation of a set of 3d mesh vertices that simulate a facial muscle group.

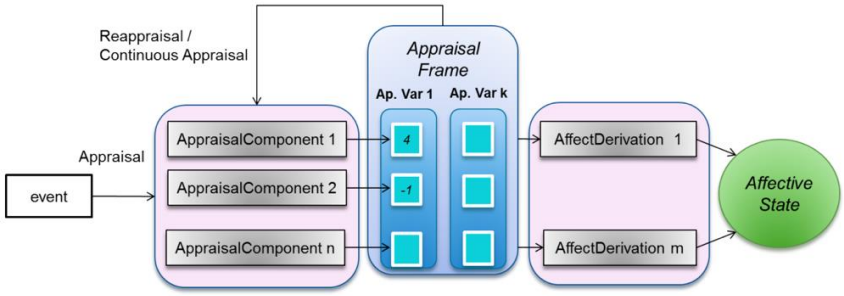

Figure 7.3: The FAtiMA Appraisal Frame Model. Process provides detail of transforming the appraisal signal into a derivation (elicitation).[201]

With annotation models summarized from Performance and Appraisal Theories, how do these two methodologies interface for the development of facial emotion corpora? Performance Theory provides a means to structure a rehearsal and video production so that targeted emotions will occur, and their facial elicitations will follow for the camera to record. Appraisal Theory informs the developer on which emotions to target and how to structure a video game scenario so that the actor can process stimuli with full appraisal and reappraisal cycles, moment by moment so that the actor and their face can be seen moving through the appraisal model path to experience the targeted emotions.

A well-designed scenario used for facial emotion expression corpus production will allow appraisal cycles to complete themselves, unlike typical dramatic or comic scenes found in live theater, films or television, that have many interrupted appraisals, as a character changes their focus from one stimulus to another. The technique of fully processing stimuli early in rehearsal and actor training is a common Performance Theory method proposed by Meisner, another Stanislavski disciple. The resulting corpus provides video clips with fully elaborated emotion processes that demonstrate the rise, peak and fall in intensity, while also giving enough data to calculate emotion label velocity within intervals of time.

As we move our discussion toward specific annotation techniques for facial emotion expression corpora, we must first consider which emotion labels should one choose to annotate, if not all of them. Just as important, we might inquire if a video game performance director should analyze each clip and annotate them one by one, as is done with general purpose facial emotion corpora used for FER systems. If we apply an Appraisal Theory approach, we will intermittently annotate meta-data tuples for each appraisal frame within a clip in the corpus. This approach would give a set of emotion label intensity values for each appraisal frame. For a human annotator to do this manually, it would be an extremely time-consuming task beyond the scope of video game workflow schedule. Instead, we see no other feasible way than to use a FER system. These software packages evolved with the development of neural networks and are a developmental cornerstone of the burgeoning industries of Affective Computing and Emotion AI. However, they do come with controversy from within the affective science communities. To fully understand the debate over the use of FERs, it is necessary to summarize the poles of the debate.

7.2.3 Emotion Essentialism Versus Constructivism

As the 21st century progresses, automated emotion recognition has become an essential part of developing autonomous emotion AI in synthetic agents. But intellectual currents directed at this technology have been at loggerheads. One side argues that emotions are dispersed but connected physiological objects generated by an evolving brain that elicit through several channels of the body using a similar neurological structure for all humans and primates. The Essentialists argue, with a salute to Darwin's late career publication, *Expression of the Emotions in Man and Animals*, that similar emotions are found universally in primates with a "basic" set forming the building blocks of all other emotions.[202] On the other side are Constructivists who argue that the brain has the proclivity to form emotion generating neurological networks that may exists in many parts of the brain, but are differentiated by experiences from the environment. As the individual develops, their family, culture and society provide the neurological basis for emotion formation through relationships and social contexts. The resulting emotion elicitations and the body parts that process affect may differ between individuals depending on how a society may reward, punish, or ignore an individual's elicitations. The Constructivists argue that at least gender, race, and age have an influence on what emotions are "allowed" to freely elicit on the most public site of the body – the face.

The impact of this debate on facial emotion expression corpora annotation is significant. The preparation of actors for facial emotion corpora production relies on performers who modulate emotions while in character. It does not take a voluminous screening of any national cinema or collection of character-driven video games to discover that the context and relationships of social events do not allow equality of emotion elicitation. Gender, race and age, if not other classifications of people, are "permitted" to elicit unequally within a society. Public

exhibition of grief for men is not judged the same as for women. Likewise, public expression of anger for men and women, nor a majority and minority race, is no less unequal. But video corpora constructed for use in training the NNs in FERs make the assumption of the Essentialists, that a pluralistic collection of faces, such as that found in the Real-world Affective Faces Database could provide enough data to train a NN to classify and recognize basic emotions on the face of persons of most any race, age or gender.[203] While the efficacy of FERs should be questioned, the basis for their possible flaws and the methodology of their use have implications for facial emotion video corpora production for training neural network facial animation controllers.

To justify the use of FERs for NPCs, flawed or not, we look at the investigatory methodology of their core research team. Ekman and Friesen took a different approach than prevailing emotion theorists of the 1960s. Rather than postulate models of emotion and then later create experiments that generate elicitation that test the validity of the model, they began by first examining elicitations (photographs) and then seeking words (folk emotion labels) that describe elicitations across multiple cultures that include seeing through Western and non-Western eyes and using labels from their respective languages. Eventually, the research moved to reassert the long-abandoned thesis of Darwin, who after documenting animal expressions in the wild and reviewing dozens of photographs from neurologist Duchenne de Boulogne, concluded that there were only six “core” expressions: *anger, fear, surprise, disgust, happiness,* and *sadness.*[204] Ekman and Friesen concurred, albeit with their own research occurring nearly 100 years later. Using English semantic labels as a trailhead to traverse into the mental systems that generate emotion, Ekman and Friesen evolved a research method that considered emotion labels as a starting point to atomize the face for an analog emotional syntax of muscle group movements

and positions. Then they attempted to connect the behaviors of those facial regions to specific parts of the brain believed to be central in the affect generating system.

Their resulting recognition taxonomy consisted of Darwin's six English words that approximated the same meanings of the words in the language of their research subjects.[205] Within a decade, they devised an annotation system called Facial Action Coding System (FACS) based on the configuration and movement of groups of facial muscles called Action Units (AUs). They looked at the muscular systems of the face and found that they work in various combinations of what are now 66 *action units* (AUs), to elicit emotions categorically. Advocates of FACS argue that a "basic" set of emotion labels correlating to positions of facial muscle groups describe families of emotion labels whose members provide more specific elicitation types within the emotion label family.[206] They also argue one can re-examine the combinations or "blends" of emotion labels using a subsequent annotation pass to derive "complex" or "compound" emotions, which are emotions that consist of combinations of the basic ones[207].

The Ekman Friesen taxonomy was deployed in an early development of a facial affect system for an NPC. In the mid-2000s, computer scientists Kozasa et al. developed a 3d head model with the ability to computationally simulate expressions of the face use the FACS taxonomy.[208] The Kozasa model was controlled by a neural network trained on fictional meta data. At that time, the researchers claimed there were no robust datasets of facial emotion corpora. Their model demonstrated affect instances for four of the six basic emotions that Ekman and Friesen identified.

Despite the appeal of the Ekman and Friesen model, not everyone agreed. Some who ascribed to variations of Appraisal Theory contended that words were unreliable markers whose meanings could change over time. The experimental psychologist Russell, a skeptic of emotion

labels and the concept of "basic emotions",[209] believed it better to map emotions on what he called a continuous *circumplex*, or unit circle inscribed within a Cartesian quadrangle.[210] The X-axis positioned an emotion on a spectrum of positive or negative feelings of *valence*, consistent with Appraisal Theory. The Y-axis showed measurements in *arousal* (intensity). An emotion label that was culturally agreeable at the time of measurement could be positioned within the circumplex if provided some tests to justify its position within a population that speaks the language of the label. Many of the OCC emotion labels and the six Ekman and Friesen emotion labels could be mapped on the circumplex as shown in Figure 7.4.

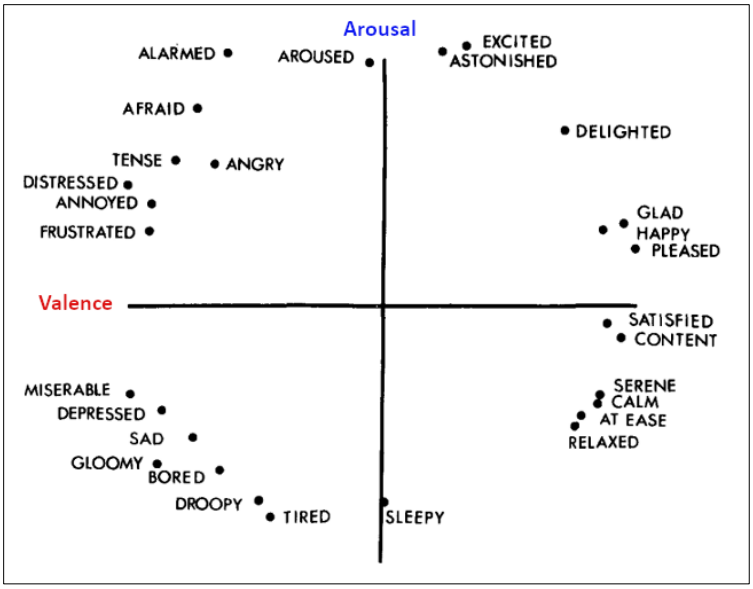

Figure 7.4: Posner and Russell's Circumplex Model. Axis labels were added.[211]

This approach reconciled for some, two theories of emotion – one that claimed universal elicitation forms as phylogenic (concepts that have later been disputed in contemporary neuroscience [212]) and another that posits the human body as having phylogenic mechanisms for emotion actuation but allows for semantic variability within cultures to define emotion labels.

Even if Ekman and Friesen, and therefore Darwin, were wrong in their Essentialist position, their emotion label concept could still be mapped on Russell's circumplex model.

For the developer of NPC facial emotion animation, the Russell model has computational prospects. We can reinterpret the Russell circumplex into a finite state machine system where each iconic emotion label is represented as the maximum value expression of that label as shown in Figure 7.5. We do this by replacing all of the emotion labels with iconic emotion expressions seen in video stills of a performing actor from their facial emotion expression corpora. A FER must be used to recognize the stills that demonstrate maximum values for each emotion label while also showing minimum values for all others. The Affect Derivation signal that exits the Appraisal Frame as seen back in Figure 7.3 becomes the activation vector that can trigger Affect Derivations in the FSM of Figure 7.5. The signal can alter the Affect State as a projection of a vector on the Circumplex. But since there are six vectors, one for each emotion label, each 100% expression of each emotion representing an iconic expression, can only be achieved if other emotions have low or no values. (Some facial elicitations are mutually exclusive.) To accommodate this restriction and to simulate multiple continuous reactions to stimuli, game engines typically implement "blend shapes" that allow affective states and their derived elicitations to be continuous blended calculations of each of the six emotion vectors for each label. Thus, the animation resulting from the vector addition of six vectors is a sequence of calculated derivations of facial deformations that can be a blend of multiple iconic emotion expressions. A representation of the circumplex implemented for a photorealistic avatar is shown in Figure 7.6.

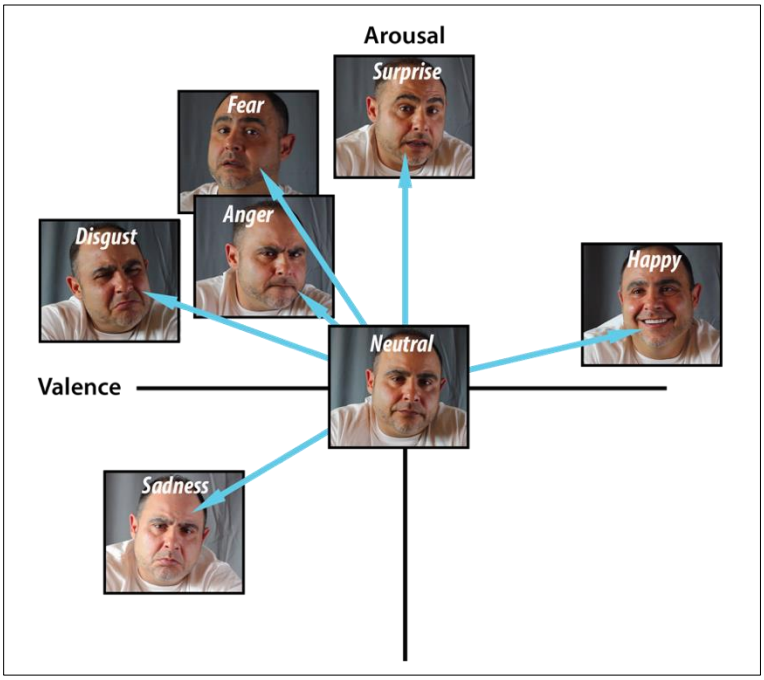

Figure 7.5: Visualizing Affect from a Corpus Actor. Locating FER analyzed images on a circumplex for optimal emotion label values of six emotions and neutral from samples in a single-actor corpus.

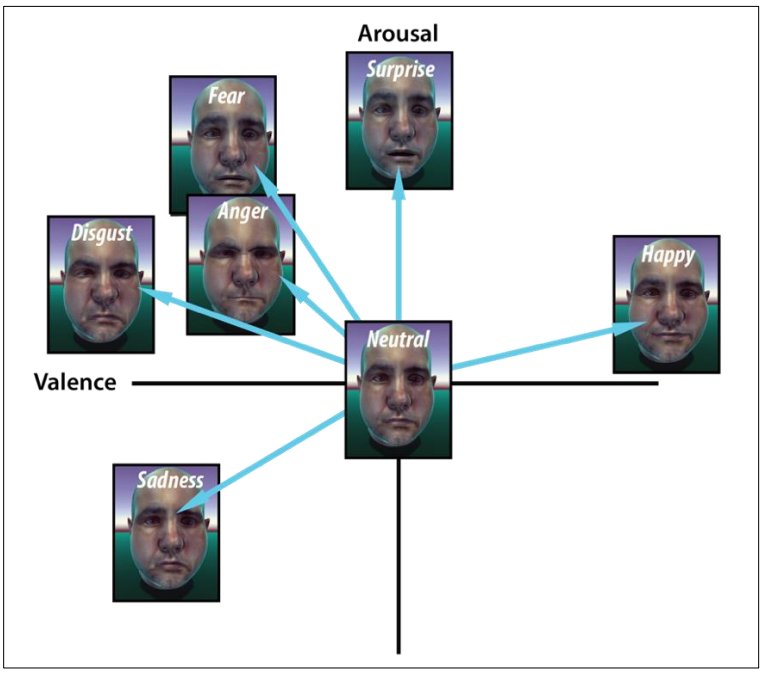

Figure 7.6: Visualizing Affect of an NPC Avatar. Locating the optimal emotion label values of an avatar on a circumplex for six emotions and neutral from a NN facial animation controller.

The emotion models presented from related work were those we have chosen to implement in part or whole. Still the polarity of the Essentialist versus Constructivist debate remains unsettled even as FERs gain market penetration in a variety of public and private spaces, and just as the public becomes wary of their use. The range of opinions on the use of FERs must be elaborated to find a methodology for autonomous animation that recognizes the limited scope of their validity and provides ethical authorship to artists whose performative behaviors are recorded as data. For experiments, we used the tool with caution, and we provided an explanation of our use followed by a justification.

7.3 Materials and Methods

The series of steps depicted below show how a NN facial animation controller is trained from a video facial emotion corpus. The steps finalize when the corpus and animation are statistically validated for their emotion label values with respect to the original actor performance emotion label values in the corpus clips. A cycle exists in the process when corpus validation yields unsatisfactory results, requiring additional corpus recording as we show in Figure 7.7 below. Annotation of emotion label values occurs at the stage of Facial Emotion Recognition. But to arrive at this step (1) a scenario for a game level must be developed, (2) an actor must be selected for the NPC emotion model, (3) their head and face must be modeled into a a 3d mesh using photogrammetric processing, (4) an off-screen actor may be needed to provide emotional stimuli, (5) a corpus of video clips must recorded, (6) facial emotion recognition annotation must be completed, (7) a neural network architecture must be designed, (8) the neural network must trained, (9) the corpus must be statistically validated for behavioral resemblance in relation to the mean emotion scores in the corpus clip recordings, and finally (10) the NPC avatar must be

validated in relation to the corpus clip recordings. This research only focuses on the annotation process that is completed during facial emotion recognition and also the process of scenario development that consist of writing and rehearsing the performance of the video clips of the corpora that will yield targeted facial emotions.

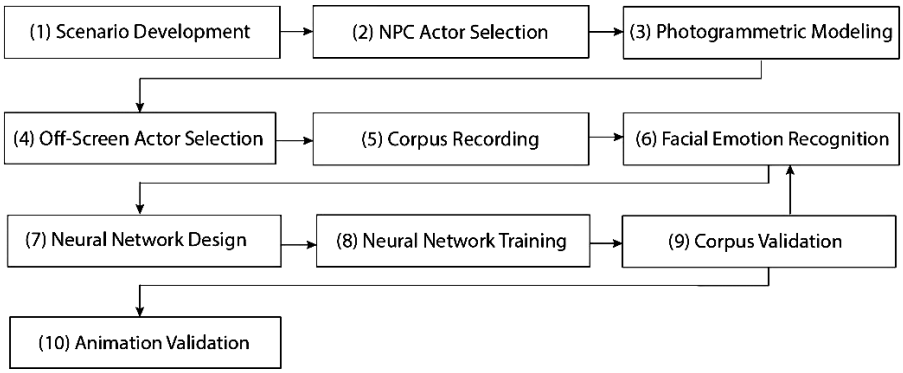

Figure 7.7: Corpus Production Workflow. Depicts steps toward creating an autonomously expressive face of an NPC using a facial animation NN controller.

7.3.1 Facial Emotion Corpus Production for Single-Actor Characterization

The annotation process of emotion labels occurs in the post-production processing stage where corpus video clips are analyzed by a FER. In our experiments, emotion label scores are intermittently processed 3 times per second. However, we should be careful to emphasize that the emotion label scores for each basic emotion are not a measurement of the actor's personal elicitation capacity, but instead that of the character. The facial emotion expression corpora must contain maximum expressions of a character's capacity to elicit some or all of six emotions (or however many emotions that a FER is set to recognize) for the events of a given game scenario, usually a level within a game. Thus, if a narrative of a game has only a use for maximizing fewer than six of the recognizable emotion labels, then there is no reason to create clips that contain maximal emotion expressions for those labels.

How then does a video game performance director produce the facial emotion expression corpus so that it contains enough of the most useful video clips to train a neural network? Answering this question is complicated by the fact that the annotation system measures emotion results. There is an obvious contradiction of interfaces when performance rehearsal processes strongly discourage asking an actor to directly give desired emotion results. Whatever an actor's training method, most actors are trained to concentrate on things in the fictional world of the scene such as objectives (goals), obstacles, and other characters in the scene, and not on the perceptions of the audience behind the invisible fourth wall. Adding to this complexity, a player within a game is often given more than one choice of actions for their avatar. These distinct choices often justify that an NPC should respond emotionally with distinct emotions processed by their appraisal of their perceptions in relation to their goals, standards and beliefs. The performance director must design a performable scenario that produces clips with emotions targeted for animation use in the game level, and with all the variability needed for all of the possible choices a player can make for their avatar. With a corpus of clips that are sufficiently varied and ripe with a character's maximal emotion instances, a corpus is ready for automatic annotation with a FER. We also propose that the FER selected use a neural network trained from a general-purpose emotion recognition corpus that has a substantial body of clips with human subjects whose ethnicity, age and gender are similar or the same as those of the actor and the NPC the actor will portray. With these considerations, we demonstrate our application of these proposed methods.

7.3.2 Dialog Behavior Tree Design and Corpora Production

The proposed approach demonstrates a technique that uses scripted performances recorded in a studio where the actor-subject performs in-character in a branching scenario with many paths.

There are two phases for this approach to facial emotion expression corpora production: (1) Designing and writing a structured scenario with paths that branch to cover all possible player choices, and (2) Recording video clips of all path performances with multiple variations of intensity.

Scenario design itself should consist of writing minimalist dialog for the on-camera character that we will call from here forward, the Emotion Model. We choose minimal dialog for the Emotion Model so that their internal emotion processing of given stimuli is less distracted by using language and memorizing a script. We also choose to capture emotion expressions with minimal dialog so that deformations of the face move from a neutral position to a maximal emotion expression of a label. This prevents the FER from misconstruing a vocalized viseme if complex dialog were spoken. The Emotion Model performs asynchronously with a scene partner that we will call from here forward, the Stimulus Source. The Stimulus Source could represent the role of the player character. Their dialog reveals more of the game level narrative. This *dyadic conversation* is mapped on a *dialog behavior tree* structure. The dialog behavior tree for the experiment is shown in Figure 7.8 beside the dialog spoken for one of the paths as seen in Listing 7.1.

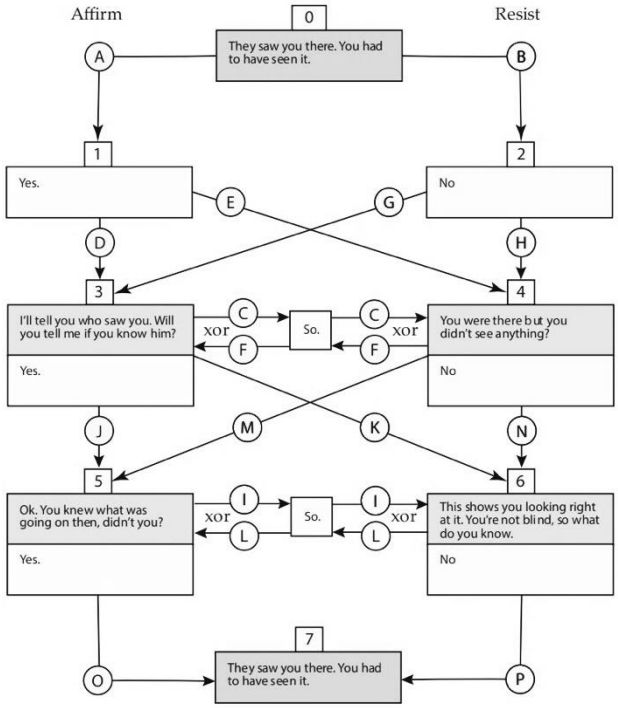

Figure 7.8: Box Nodes at Dialog Turns (3, 4, 5, 6) and Monolog Events (0, 1, 2, 7). Edges represent mental actions. As an acyclic graph, a path can use edge C xor F and I xor L.

LISTING 7.1: SAMPLE SCRIPT GENERATED FROM DIALOG-BEHAVIOR TREE

```
STIMULUS SOURCE (0): They saw you
there. You had to have seen it.

EMOTION MODEL (2): No.

STIMULUS SOURCE (4): You were there
but didn't see anything?

EMOTION MODEL (4): Yes.

STIMULUS SOURCE (5): Ok. You knew
what was going on then, didn't you?

EMOTION MODEL (5): So.

STIMULUS SOURCE (6): This shows you
looking right at it. You're not
blind. So what do you know?

EMOTION MODEL (6): No.

STIMULUS SOURCE (7): They saw you
there. You had to have seen it.
```

Listing 7.1: Script Generated from Dialog-Behavior Tree. Depicted in Figure 9. Script shows the path of the conversation through nodes 0-2-4-5-6-7 along edges B-H-M-I-P.

Role preparation for the actor need not differ much than for any role for theater or screen media except that the process of recording actual dialog is not the focus for facial emotion expresion corpora. Design of a dialog behavior tree for our experiment consists of distinct and fixed start and end nodes. The Emotional Model actor that will play the NPC is the one whose emotion labels the neural network will train and simulate with responsive animation. The Stimulus Source provides input stimuli to trigger the Emotion Model to elicit. Animation in the form of predicted emotion label values in response to game scenario stimuli is the result of the neural network receiving stimuli as data in form of appraisal variable values, including behaviors of the player,

and other agents, events and object in the game level. The responsive neural network embedded in the NPC sends its response signal in the form of normalized floating point numbers that move actuators in the 3d facial mesh from a zero (neutral) to 1 (fully articulated maximum emotion label). The response signal is what is depicted back in figure 3 to trigger an Affect Derivation and a change in Emotion State.

For the scenario presented in Figure 9, the enumerated box nodes (0-7) are dialog turns between the Stimulus Source (gray boxed dialog) and the Emotion Model (white boxed dialog). The left- and right-hand side represent distinctly opposing actions (*affirm* versus *resist*) for the Emotion Model. These are the actions that must be given to the actor for rehearsal. As the scene progresses from the start node 0 to the end node 7, we anticipate two emotion labels to have the highest intensities: *anger*, and *sadness*. To achieve this goal, we collaborate with an actor to generate a table of data using Hagen's six questions. The questions and answers are presented below in Table 7.1.

TABLE 7.1: HAGEN'S SIX QUESTIONS WITH ANSWERS FOR A DIALOG BEHAVIOR TREE

Who am I?	Data	Role (noun): *mechanic, gambler, bachelor*
	Character Narrative	I am a *mechanic*, and a *gambler*. I'm a *bachelor*. I like playing cards for money without anyone judging what I am doing.
What are the circumstances?	Data	Events (noun): *interrogation, crime*
	Character Narrative	I was brought into the police office for questioning because someone upstairs in the house where I was gambling was found dead.
What are the consequences?	Data	Future States of Being: *freedom, incarceration, injury, death*
	Character Narrative	I might have seen something, but I owe money to the house. If say too much, I could bring trouble to myself. If I satisfy the detective, I could get freedom. If I say too much, the house bookies could hurt me. If I say the wrong thing, I could be incarcerated.
What are my relationships?	Data	Other Roles (nouns): *detective, gambling partners, work friends, mother, brother*
	Character Narrative	I don't know this *detective*. I barely know the *gambling partners* I play cards with. I have *work friends* who know me. My *mother* and *brother* live far away.
What do I want?	Data	Scene Objectives (nouns): *handshake, open door, trust*
	Character Narrative	If the detective gives me a *handshake*, he will offer *an open door* to go home. To get that, I will gain his *trust* that what I am saying is the truth.

What is my obstacle?	Data	Obstacles (nouns): *officers, guns, sight line, detective, evidence envelope, questions*
	Character Narrative	There are *officers* with *guns* who won't let me freely walk out. A locked door also blocks my way. The *detective* presents *evidence* from an envelope that creates *questions* I do not want to answer.
What do I do to get what I want?	Data	Action (verbs): affirm, resist
	Character Narrative	I will cautiously *affirm* the facts they present unless something could get me in trouble. I must *resist* questions I am afraid to answer, that could cause the gambling house to seek retribution for providing evidence against them.

The Stimulus Source triggers the Emotion Model to process accusation, coercion and negotiation with three possible answers: "Yes", "So" and "No." If "Yes" was chosen, the path continues straight down the "Affirm" action path triggering more appropriate responses from the Stimulus Source. If "No" was chosen, the path is directed toward a "Resist" action path and the Stimulus Source answers accordingly. Some paths allow for crossing over with contrary responses. A "No" on the "Affirm" side will direct the path down to the next level and to the opposite side into the "Resist" side. Vice-versa is also true. A "So" answer will also cross over but does so while retaining the same level in the tree. We choose that the graph be acyclic so that the performance does not have any internal cycles and to create a consistent data set of equal lengths for each path.

Letters in circles identify each edge for use when identifying regions of paths recorded as video clips for emotion analysis between dialog turn nodes. Edges represent the transition between dialog turns. Emotion elicitations as reactions to stimuli typically occur in performance at the beginning of an edge immediately after hearing an opposing character speak. Emotional elicitations as actions of agency (those facial expressions that use the face as a directed action) typically occur at the end of an edge. An example of an action elicitation of agency would be

when a person *consciously* demonstrates *disgust* on the face to physicalize the action (verb) of *disapprove* or *reject*.

Using the rules of an acyclic graph, with a single start and dinstinct end node, there are 32 paths through the tree. Each path was recorded 9 times with three distinct degrees of intensity of action: high, medium and low as suggested by Wingenbach et al. for the production of the Amsterdam Dynamic Facial Expression Set.[213] For each path, the three groups of clips (9 total) represent the variations of need of the Emotion Model's objective during the scripted game scenario. The Emotion Model performed in precise synchronization for each clip with a corresponding pre-recorded clip of the Stimulus Model for each of the 32 paths. The production setup is shown in Figure 7.9. Thus 9 clips times 32 paths yielded 288 total clips of the Emotion Model in the corpus. With the corpus complete, its clips were used for training our neural network.

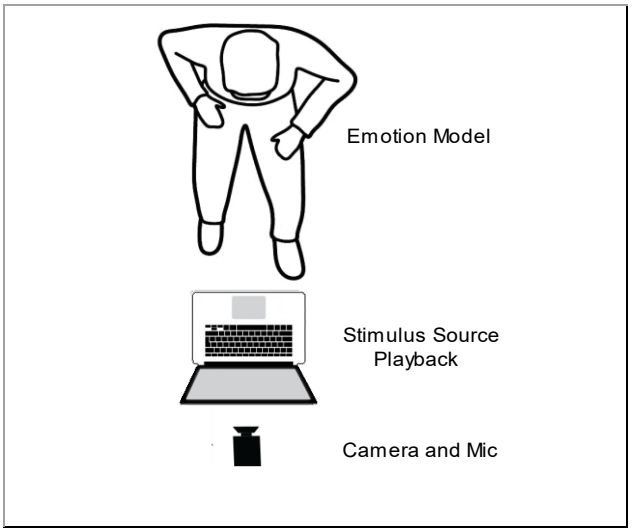

Figure 7.9: Production Setup. Emotion Model sample recording for facial emotion expression corpus.

7.4 Discussion

The preliminary experiments we conducted demonstrate statistical resemblance in emotion intensity and emotion velocity. The results suggest that our proposed methods are efficacious.

There are several important issues to consider before undertaking further experiments. The use of FERs needs cautious consideration. Their persistent use in scrutinizing human behavior presents some problems of accuracy, and some of the theoretical underpinnings that justify their use continue to be eroded by recent neuroscientific research.

7.4.1 The Emergence of FERs and Doubts About Their Use

The influence of Ekman's and Friesen's influence cannot be underestimated. When their concept of the six basic emotions was published in 1971[214], it defied a trending belief in the social sciences argued by anthropologist Mead and others, to give more import to cultural determinants of emotion expression.[215] In the early 21st century, when facial emotion corpora were first being assembled for the purpose of designing neural networks that could demonstrate fundamental sentient recognition, the Ekman and Friesen definition of basic emotion labels combined with Russel's Circumplex model, took hold not only in the imagination of academics in the study of affect, but in the hopes of investors in several FER systems that are now at the time of this writing, multimillion dollar corporations in the field of developing commercial FERs. One can see in the interfaces of FERs the design influence of the research dating from 1960 through to 2005 by looking at Figure 7.10, a commercial interface design for Noldus FaceReader 8 circa 2021. The Cartesian plane and the unit circle provide mappable positions for affect based on valence and arousal readings. The emotion labels are positioned at specified coordinates determined by data-driven cultural research and by the statistical culling of expression recognitions completed by annotators of the corpus used to train the NN of the product. Other FERs show similar designs and/or uses of the same six emotions.

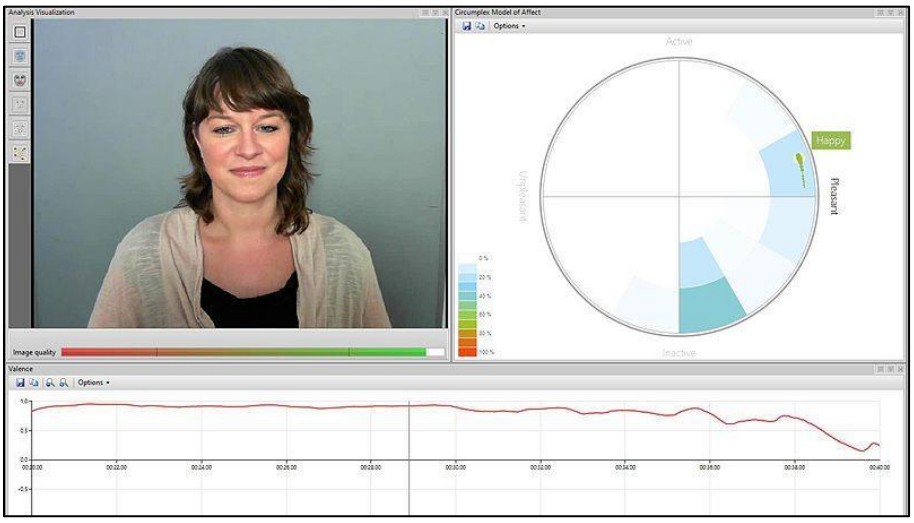

Figure 7.10. Noldus FaceReader Real-Time Expression Analysis Interface. The current frame of motion video shows its coordinate reading of the "happy" label.

Despite the general acceptance of Ekman and Friesen's semantic divisions of emotion and the Essentialist viewpoint that it decries, disagreement from Constructivist emotion theorists has been joined from multiple fields within the social and biological sciences, including psychology and neuroscience. A reading of Ekman's publications finds among the collection many defensive arguments against alternative theories of emotion identification, arguing against neuroscientists like Feldman Barrett who assert that learned "regulation" of emotion elicitation stems from cultural roles. Feldman Barrett, among others, find that societal penalties are imposed for authentic but "inappropriate" elicitation at socially important events and places[216]. She also argues that the perception of emotions is affected by relationships between persons of different identities, such as gender.[217] In consort, neuroscientist Damasio argues contrary to Ekman that the neurology of emotions is dispersed throughout the brain and cannot be seen as functioning with emotion-specific regions, a concept that Essentialists portend. Neuroscientists Immordino-Yang et al. hold that somatosensory mechanisms in the brain are subject to disparate influences within the body by both culture and evolution, thus altering the awareness of emotions

themselves, and their elicitations.[218] Awareness of emotion in oneself and others is a fundamental premise of emotion Essentialism. Semantic labels and the apparent facial expressions correlated, Feldman-Barrett contends, is what links these neuro-biological activities together, and not any specific region in the brain holding a primordial commonality of expression and recognition.[219]

Ekman has also defended his model against those that argue against basic emotions[220], which is also at the core of his research framework and is the basis for FERs. While the evidence he has presented throughout his career provides an empirical construction of the identification of emotions by semantic labels after their elicitation, there is not agreement on the specific neurobiology of emotion generation and emotion specific genesis and elicitation. And thus, any system of annotation that uses human language as its primary identifying signifier to differentiate emotions is subject to variability of semantics or scientific discovery that occurs over time by changing the meaning of words that describe emotion.[221] Even Ekman later conceded that there are more than six basic emotions and probably more to come, though he vigorously counters that any basic emotion must pass a set of 13 criteria that qualify an emotion's individuation, impulsiveness, commonality among primates, brevity of feeling, and implications of moral outcome.[222] He anticipates new entries from other languages, suggesting a more panoramic sense of label etymology. Despite a ruckus of contradictory discourse, the emotion recognition software industry integrated an Essentialist framework to develop substantial products during the last decade. Annotation of facial emotion expression corpora has had a bountiful decade with nearly a dozen widely used annotated facial emotion video corpora available for research and commercial applications such as FERs.[223]

Handing off the human process of recognizing emotion to AI has invited critical inquiry and skepticism in the popular culture at large. Much of the doubt cast on facial emotion recognition comes from those more persuaded by emotion Constructivism, fearing the potentially deleterious effects of misconstruing emotion in a variety of contexts. Before embarking on a disclosure of how to use existing FERs to create single-actor corpora for facial emotion controllers of NPCs in video games, it is worth considering the assumptions, production methods and shortcomings of FERs. With an understanding of their limitations, the artist and developer can optimize these limitations for maximum expressivity of fictional character design.

Training neural networks for facial emotion recognition requires large sets of video corpora drawn from populations in specific regions of the world that a FER intends to recognize. Each annotator must recognize the ethnically specific variation of each of the 66 action units of the face. Then, they must recognize them in combination as they elicit the basic emotions. This intimate human judgment is necessary to provide meta-data for each video clip within a corpus used for training NNs for emotion recognition and for subsequent NN validation required after initial training. It is during the validation stage that NN classifications are statistically compared with human annotator classifications of the same facial images. This later stage is what is publicized as recognition accuracy. For a NN to be sufficiently trained, it must be sufficiently accurate for recognizing the basic emotions on faces for at least the age, ethnicity and gender demographics of the corpus used to train it. Following the Essentialist viewpoint that there exists “ground truth” definitions of emotions that can be detected in the human face, most facial emotion recognition NNs use supervised learning methods. This approach imposes pre-defined facial expressions in the training algorithm for the neural network to classify distinct emotion labels within a video corpus. Usually, those pre-defined expressions are representations of the six

"basic" emotion labels. We will identify some of the evidence that fuel doubts about FER accuracy and the validity basic emotions so that the game developer and corpora designer can frame their own uses and discussion of annotation with some epistemic precision.

Emotions are not themselves elicitations of the face, nor are any other neuro-muscular or physiological behavior. Neuroscientist Damasio concludes that emotions are an unconscious combination of sensory awareness of changes to the body caused by external or imagined stimuli that activate previously experienced neural patterns stored in the brain.[224] To paraphrase his discussion, an emotion is a feeling about a collection of feelings. Those previous neural patterns are the result of lived experiences (often during childhood) that are shaped by their contexts. These contexts include the physical reactions and perceptions of others – perceptions explicated by human language. Feldman Barrett provides a model as to how language-specific emotion labels make for rapid emotion recognition. Her research shows that these semantic constructs are based on awareness of the context of the subject and the observer, and that they can be artificially affiliated with "stereotypical" facial forms of emotion elicitation.[225] When an eliciting person was presented in a photograph of only their face, an annotator provided an emotion label. But after revealing the full context of the elicitation, the experiment indicated that the "true" emotion felt and observed in person differed from the emotion label that was originally perceived by the observer of the photograph. Additionally, when observers were given more than six emotion labels to describe a corpus subject, the accuracy of human recognition declined. Feldman Barrett's conclusion was that emotion perception is a process of fitting the stimuli of the face to fit the knowledge available to the observer. Some knowledge is empirical, the information of as much of the context that a viewer can perceive. Often, the viewer "recognizes" an emotion without awareness of all relevant visible facts around the face. In Feldman Barrett's

experiments, a photograph of an isolated face provoked from a respondent a different emotion when the face was shown in the context of an event in progress with other persons and a causal incident in the frame.

Contextual knowledge also includes the linguistic and semantic constraints of an observer's language and their idiosyncratic memories of real people responding to remembered events. Media representation of events portrayed in illustrations, photographs, and videos also can supplant an observer's association with an emotion label and the facial expression they correlate. Language in the form of captions, commentaries and dialogues combine to shape an emotional interpretation of a memory or media representation. Even awareness of the existence or intensity of an emotion may not be universal, according to one neuroscientific experiment. In experiments conducted by Immordino-Yang et al., culturally influenced degrees of permitted expressiveness defined not only language used to describe emotions, but the very awareness that some of the feelings associated with emotion-generating events even exist.[226]

As the Essentialist argument for the existence of universal basic emotions across all human species is tested by neuroscientists, we find data produced in some transcultural experiments undermine the precision of the Essentialist premise, though they do not contradict it entirely. One study developed their own FER to find degrees of variance across race, class, gender and age of the emotion *happiness*. Fan et al. chose to focus on the set of action units that recognize only the much studied "Duchenne smile," an elicitation that requires significant deformation of the mouth, lips and cheeks[227]. It is identified as the most authentic elicitation of genuine happiness by Ekman.[228] The study found a 12% difference in intensity across distinct races found in the Real-world Affective Faces Database. A 12% difference in expression for the Duchenne smile, one of the most exemplary of elicitations, allows us to consider what Damasio

and other neuroscientists proposed – that while the brain has the propensity to develop channels that classify emotions, experiences framed by culture and language may play a role in shaping the channel and the form of the elicitation. Thus, the accuracy of FERs across demographic distinctions may vary substantially for some emotion labels.

7.4.2 Faciasemiotic Applications of FERs

While we recognize that Feldman Barrett's position undermines the validity of using FERs for detection of psychological states, it does validate the use of FERs as a tool for identifying iconic expressions. Such "iconic" expressions can serve as distinctive and well-formed signifiers that communicate emotional ideas, and that signify a "faciasemiotics" that operates in stage and screen media.[229] Video game character design is a form of creative fiction, not unlike producing theater, film or television. Characterization may be shaped by actors, but as many theater directors, filmmakers and scholars of the media and performing arts have demonstrated and dissected, much of the perception of emotional meaning consists of a viewer's or player's preconstructed ideas about what causes emotions in people, how those emotions are elicited, and how the context of observed behavior effects the judgment of a character's emotional state. These preconceived notions comprise the signified of any emotion label where the sign is the facial expression. Cognitive film scholar Smith asserts that films prepare the audience for a character's perceived emotional disposition through the use of cued emotion markers.[230] Schiller, who studied Ekman's research as a form of facial semiotic taxonomy, would frame Smith's cueing as conjuring the signified of an emotion in the viewer's mind.

Smith elaborates on how emotion markers are presented through the actor's face as they appraise the stimuli of a scene, while others are presented by the film's lighting, camera position, props and setting – the mise-en-scene.[231] In one extensive example, Smith discusses how in the

film *Stella Dallas*, the viewer's perception of Stella's emotional state is presented as a dilemma of judgment that gradually reveal its truth. There are events that occur in the film that rely on a viewer's pre-existing memories of "appropriate" behavior for Stella's role that anticipate what Stella "should" be feeling – what Smith calls "microscripts" – that contradict what viewers see in the film of what she is actually feeling. The actuality of Stella's feeling is delivered through closeup shots of Stella's face internalizing the events around her at moments when other characters cannot see her face, but the viewer can. Stella's face changes to expressions of "expected" emotions when other characters do see her face. Once the actual feelings are presented to the viewer through the face, then the authorially selected emotional cues that truly represent Stella's state follow to reinforce the recognition of the privileged viewer perspective. The facial elicitations we observe on Stella consist of what Smith calls "emotion prototypes". These are recognizable facial elicitation forms that allow the viewer to quickly assign meaning based on remembered iconic representations of faces with a correlating emotion label. For the cinematic telling of the Stella Dallas film narrative, the use of facial emotion prototypes gives clear motivation to the viewer for Stella's behaviors. Some behaviors, including facial expressions, intend to deceive other characters, and others are authentic elicitation. The viewer eventually realizes which are true and which are false because they see contradiction and context. Both the truthful and false elicitations are recognizable to the viewer because they are prototypes, with less ambiguity.

Previously, when presenting the views of emotion Constructivist, the terms "stereotype" and "iconic" were used to describe definitive images of an emotion label. By inference, one would assume that since FERs declare their validity at recognition with an Essentialist premise, the resulting classifications would also be "stereotypical" or "iconic." It should be clear that in

the context of scrutinizing the accuracy of FERs, this word has been used appropriately because it suggests that the images used reduce the complex and complete meaningfulness of the experience of the subject to a single word. However, for the purpose of creating facial emotion corpora, an actor can intentionally create instances within a corpus of studio-produced video clips that include emotional prototype images for the character, much as Smith uses in his discussion of the emotion system in cinema. Instead of reducing the meaning of a facial image to a word, a single-actor facial emotion expression can provide a customized lexical system for NPC elicitation that has a character's idiosyncratic emotion prototypes in the form of facial emotion images for each basic emotion label, and many variant blended combinations. We find evidence of this in the experiments of psychologists Berry and Brown. They contend that actors use contrastive exaggerations of facial expressions to clarify and differentiate their characters in consort with one another.[232]

Given that FERs generally only score 6 emotion labels (some analyze *contempt* and *neutral* in addition to the basic six), is that enough differentiation to train a neural network in an NPC to sufficiently resemble the behavior of the actor's character? This issue remains to be seen. We have produced two experiments that statistically analyze and compare the facial expressions of performing actor of facial emotion video corpus with the facial animation behavior of their avatar trained from those same clips. One experiment compared emotion label intensity[233] and the other studied emotion label velocity,[234] both for targeted emotions. The results are promising and demonstrate the viability of the methodology proposed. We encourage the game production and research communities to continue experimentation in this domain of game development.

As the research of this chapter is an investigation in how to use a scientific tool for an artistic process, we accept the limitation of the few emotion labels and their scrutinized accuracy.

We attempt to expand the application of the tool beyond its intended use. One strategy to create new *derived emotion labels* is to develop an application for analysis on the FER-generated data of the six basic emotions that examines *proportional combinations of the six basic emotions.* Following Ekman's concept that emotions can be blended, and that there are families of emotions,[235] a scenario, rehearsal and clip production workflow would have to be designed for the actor as a character to elicit these complex emotions. Then, proceeding with the idea that a complex emotion consists of idiosyncratic percentages of the basic emotions, the definition of a complex derived emotion would have its own distinct label and would then be defined by a range of specific proportions of the basic emotions as reported by the FER.

7.5 Conclusion

We recognize that our approach to annotation for NPC behavior presents only a part of the potential for analyzing social behaviors with computational means. But there are also uses for our approach for modeling other forms of synthetic agents, such as conversational agents in the workplace, healthcare provider agents in medicine, and instructor agents in education. By combining research in psychology and the performing arts, we believe researchers can imagine a wider range of motivations and responses in modeling human behavior through emotion AI in synthetic agents. Annotation methodology is a seminal step in the behavioral design of a neural network that informs emotion AI. For single-actor facial emotion corpora, annotation for their video samples can be systematic and targeted at the emotion labels needed for the application. Before any annotation approach is standardized for creating facial emotion expression corpora, continued research and experimentation must ensue toward developing methods of validation of resemblance between the synthetic agent such as an NPC, and the actor that it represents.

8 CONCLUSION AND FUTURE WORK

We have established the viability of the workflow and NN architecture using statistical evidence. We considered the accuracy of our NN determined that within a precision range of $1/100^{th}$ our NN architecture produced a mean squared error score below the precision range (return to table 3.1). But since NN accuracy only looks at overall without regarding to the temporal position of any data point in time, we reasoned that consecutive time-series data should use another method to validate our NN architecture and corpora that trained it. After sorting the data set by common segments of video clips that traverse the same edge of a path through the dialog behavior tree, we proved at minimum a probability of resemblance of facial animation behavior within ±1 standard deviation for targeted emotion labels in relation to the expression of the actor of a corpus. Both intensity and velocity were analyzed and for targeted emotions, our statistical analysis demonstrates significant resemblance between the NN-controlled photorealistic avatar and the actor used to train the NN.

But several questions remain unanswered that would require additional experimentation. First, is the dialog behavior tree too sparse? Given that a complete set of clips for 9 takes for each path generated 288 clips totaling 3.6 hours of footage for each corpus, the scale was appropriate for our small study. Would a larger study with more clips demonstrate significantly different results? Our production parameters required 8 hours to shoot all the video for each corpus. Four actors were used for the experiments. While this may seem like a large quantity of video, a commercial video title often uses hundreds of hours of reference and scratch audio video for an hour game play. More clips would provide more samples and a more extensive tree structure would provide more variations from which to train the NN. More samples may improve performance for all of the measurable parameters demonstrated in our study.

Also, we only worked with one NN architecture for all 6 emotions and neutral, despite using separate NNs for each emotion label. Further experimentation may lead to a more optimized architecture that could improve resemblance for each emotion label. Elicitations that are recognized as surprise or fear occur very quickly and are often sustained for long periods of time. An architecture that optimized temporal data with more LSTM cells combined with a more frequent frame rate of emotion analysis (greater than 3 per second), could produce more accurate data for those targeted emotions. After careful analysis of the muscular dynamics, it would be a worthy direction to customize and optimize unique NNs for each emotion label.

We also believe that it would be useful to experiment with augmentation of a corpus so that if one emotion label is showing low resemblance, we could design an experiment where new clips could be added to the corpus to allow for retraining the original NN to improve resemblant behavior for more than one emotion label. The process of retraining should be a whole new direction that would allow for incremental improvements with more clip production over time.

Among the corpora that we produced but did not complete the process of NN training, was for two datasets involving two actors that appeared in a film. Our intention was to use a combination of corpus video production clips with footage from the film to create an avatar with NN controlled facial animation that could be used to create a shot in the film. We suspended this approach for the set of chapters of this dissertation because our experiments only address animation of the neuromuscular flesh of the face. Complete head animation was deemed necessary. However, the face consists of other movement systems for which we have no system developed to deploy, such as the eyes and eye lids and the effect of facial orientation as emotion elicitation effects the neck and its effect on facial orientation along six degrees of freedom. Nonetheless, a future study should focus on simulating living actors that appear in films. A

successful result could provide a security model that would reduce the risk of a halted production in the event of actors being injured during production.

It should be mentioned that the experiment relied on off-the-shelf tools that are also evolving. Relying on FERs for automatic annotation is a necessary tool to make the workflow viable, though we realize we are dependent on each FERs quality and diversity of corpus subjects. But FERs have much to improve to allow for the most diverse collection of human beings to participate in this technology. It would be prudent to cross-validate our FER data with two or more FERs rather than rely on only one. FERs are very expensive to license on an annual basis. For this study, two different systems were used with license fees discounted by 80%.

Each one of these future experiments that we have highlighted in this chapter must ultimately lead to NPC animation that creates an experience of the animation that simulates an actor's character design. Several important experiments would be essential at proving the viability of this research. First, the NPC animation itself needs to be synchronized to the Stimulus Source and fed to a FER to compare the generated data of the animation with a mean-of-means for each frame of each take of the Emotion Model. An RMSE score for sequential time-series data for frames synchronized to the Stimulus Source could be persuasive. However, one must keep in mind that the NN should never behave exactly like any one video clip of an actor's performance of a scene or even the mean of means from many takes of the actor's performance. The emotion label values should fall within an acceptable range much as we might expect an actor to perform a character in a production of a play, a little differently each night. We do not yet have an RMSE score range or other measurable method for this test to validate resemblance.

Other methods that could be considered would fall into the category of measuring audience or player response. While audience or player response analysis could be very useful, it is somewhat out of the domain of Computer Science, and would be best conducted in conjunction with media and psychology scholars.

Even with a rich list of future directions for research, each experiment we conducted for this research led to many possible parameter changes in the experiment that would require additional experimentation on multiple tributary paths of discovery. As is always the case, one must make boundaries and cull the results for publication. There remains an abundance of new experiments to complete that will perfect the workflow and the software system that we developed. Upon completing more tests and implementing the approach in a publicly available game, only then would we discover if video game players will engage with similar enthusiasm to an autonomously animated face of an NPC as they do with the same character animated by traditional motion capture and hand-crafted animation techniques.

We have no doubt that in the next few years, the current trend in AI research will demonstrate autonomous facial elicitation much as we are currently experiencing highly persuasive autonomous writing and visualization by AI that simulated living writers and graphic artists. We believe the current trend in using extent works of literary and visual art as free feed for training algorithms will lead to litigation and possibly legislation to protect the intellectual property of working and deceased artists. No matter the outcome, this trend will eventually consume works in the performing arts. One way to counter the tendency of positioning the performing artists as an exploitable content provider is to augment existing art forms to include corpora production as a new medium, and to adapt techniques of acting to optimize corpora production. It is our hope that this research will provide a path or an informed model for future

technology development designed to help artists adapt to the economies of media production, dissemination and consumption.

REFERENCES

[1] Calleja, G. (2009). Experiential narrative in game environments in breaking new ground: innovation in games, play, practice and theory. In Proceedings of DiGRA 2009.

[2] Ekman, P. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press: New York.

[3] de Maupassant, G. (2007). The short stories of Guy de Maupassant. Black's Readers Service: Roslyn, NY.

[4] Braunberger, P., (Producer), & Renoir,J. (Director). (1936). A day in the country [Motion picture]. France: Panthéon Productions.

[5] Chatman, S. (1980). What novels can do that films can't (and vice versa). Critical Inquiry, 7(1), 121-140.

[6] Smith, G. M. (2003). Film structure and the emotion system. Cambridge University Press, 41-64.

[7] Margolin, U. (1986). The doer and the deed: Action as a basis for characterization in narrative. Poetics Today, 7(2), 205-225.

[8] Charles, David (2014) Aristotle on agency, The Oxford handbook of topics in philosophy (online edition, Oxford Academic)

[9] Murray, J. H. (1997). Hamlet on the holodeck: The future of narrative in cyberspace, MIT Press: Cambridge.

[10] Maes, Patti (1997) Intelligent software, In IUI '97 Proceedings of the 2nd International Conference on Intelligent User Interfaces, 41-43.

[11] Rickel, J., & Johnson, W. L. (1997). Integrating pedagogical capabilities in a virtual environment agent. In Proceedings of the First International Conference on Autonomous Agents (pp. 30-38).

[12] Lane, H. C., Cahill, C., Foutz, S., Auerbach, D., Noren, D., Lussenhop & C., Swartout, W. (2013). The effects of a pedagogical agent for informal science education on learner behaviors and self-efficacy. In Lane H. C., Yacef, K., Mostow J. Pavlik P. (eds) Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science, vol. 7926. Springer: Berlin, Heidelberg.

[13] Nixon, M., Pasquier, P., & El-Nasr, M. S. (2010). DelsArtMap: Applying Delsarte's aesthetic system to virtual agents. In International Conference on Intelligent Virtual Agents (pp. 139-145). Springer: Berlin, Heidelberg.

[14] Maes, Patti (1997) Intelligent software, In IUI '97 Proceedings of the 2nd International Conference on Intelligent User Interfaces, 41-43.

[15] Turing, Alan (1950). Computing machinery and intelligence. Mind, Volume LIX, Issue 236, October 1950, 433–460.

[16] Turing, Alan (1950).

[17] Descartes, René (2006). A discourse on the method of correctly conducting one's reason and seeking truth in the sciences, trans. MacLean, I., New York: Oxford University Press, 46.

[18] Descartes, R. (2006).

[19] Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. Brain Research Reviews, 26(2-3), 83-86.

[20] Johnson-Laird, P. N., & Oatley, K. (1998). Basic emotions, rationality, and folk theory. In Consciousness and Emotion in Cognitive Science (pp. 289-311). Routledge.

[21] Picard, R. W., & Rosalind, W. (2000). Toward agents that recognize emotion. VIVEK-BOMBAY, 13(1), 3-13.

[22] Price, R. H. and Bouffard, D. L. (1974). Behavioral appropriateness and situational constraint as dimensions of social behavior. Journal of Personality and Social Psychology, 30(4), 579-586.

[23] Yin, M., & Sun, Y. (2015). Human behavior models for virtual agents in repeated decision making under uncertainty. In Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).

[24] Gratch, J., Okhmatovskaia, A., & Duncan, S. (2006, November). Virtual humans for the study of rapport in cross cultural settings. In 25th Army Science Conference, Orlando, FL (pp. 27-30).

[25] Zhao R., Papangelis A. & Cassell J. (2014). Towards a dyadic computational model of rapport management for human-virtual agent interaction. Bickmore T., Marsella S., Sidner C. (eds.) Intelligent Virtual Agents. IVA 2014. Lecture Notes in Computer Science, vol 8637. Springer: Cham.

[26] Picard, R. W., & Rosalind, W. (2000). Toward agents that recognize emotion. VIVEK-BOMBAY, 13(1), 3-13.

[27] Bickmore, T., Bukhari, L., Vardoulakis, L., Paasche-Orlow, M. & Shanahan, C. (2012). Hospital buddy: a persistent emotional support companion agent for hospital patients. 492-495.

[28] Dormehl, L. (2018), A.I. border agents could use machine smarts to tell if travelers are lying, Digital Trends, 18 May 2018, accessed on 2 March 2023 at https://www.digitaltrends.com/cool-tech/ai-border-airport-virtual-agent/

[29] Deahl, D. (2018), The EU plans to test an AI lie detector at border points, The Verge, 31 October 2018, accessed 2 March 2023 at https://www.theverge.com/2018/10/31/18049906/eu-artificial-intelligence-ai-lie-detector-border-points-immigration

[30] Lucas, G. (Director) (1980) The empire strikes back [FILM], Lucasfilm Ltd., 20th Century Fox.

[31] Trull, T. J., & Widiger, T. A. (2013). Dimensional models of personality: the five-factor model and the DSM-5. Dialogues in Clinical Neuroscience, 15(2), 135.

[32] Malatesta, L., Raouzaiou, A., Karpouzis, K., & Kollias, S. (2009). Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. Applied Intelligence, 30(1), 58-64.

[33] Morgans, Julian (2018), How video games cast actors just like movies do, Vice, 10 April 2018, accessed on 2 March 2023 at https://www.vice.com/en_uk/article/gymde7/how-video-games-cast-actors-just-like-movies-do

[34] Stuart, K. (2016). Video games where people matter? The strange future of emotional AI. In The Guardian, 12 October 2016, accessed 2 March 2023 at https://www.theguardian.com/technology/2016/oct/12/video-game-characters-emotional-ai-developers

[35] Torres, S. (2014), Video game characters modeled after real people. In Venture Beat, 28 June 2014, accessed on 21 September at https://venturebeat.com/2014/06/28/video-game-characters-modeled-after-real-people/

[36] McNulty, Thomas (2020), A screen is a screen: Actors are moving to video games as a new medium for their career, In Hollywood Insider, 25 February 2020, last accessed on 2 March 2023 at https://www.hollywoodinsider.com/video-games-actors/

[37] Hart, H. (2005). Do You See What I See? The impact of Delsarte on silent film acting. Mime Journal, 23(1), 184-199

[38] Quantic Dream, Beyond two souls [Video game]. (2013). Sony Interactive Entertainment.

[39] Naughty Dog, The last of us [Video Game]. (2013). Sony Interactive Entertainment.

[40] Lutterbie, John (2011), Towards a general theory of acting: Cognitive science and performance, Palgrave Macmillan: New York, 103-130.

[41] Smith, G. M. (2003), 41-64.

[42] Wiley, N. (2003). Emotion and film theory. In Studies in Symbolic Interaction, Vol. 26, 169-187.

[43] Ortony, A., Clore, G. L., and Collins, A. (1990). The cognitive structure of emotions. Cambridge University Press: Cambridge.

[44] Lazarus, R. S. (1991). Cognition and motivation in emotion. American Psychologist, 46(4), 352.

[45] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161–1178.

[46] Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. Emotion, 11(4), 705.

[47] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4), 261-292.

[48] Malatesta, L., Raouzaiou, A., Karpouzis, K., and Kollias, S. (2009). Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. Applied Intelligence, 30(1), 58-64.

[49] Lisetti, C., and Hudlicka, E. (2015). Why and how to build emotion-based agent architectures. In R. A. Calvo, S. K. D' Mello, J. Gratch, and A. Kappas (Eds.), Oxford Library of Psychology. The Oxford Handbook of Affective Computing (pp. 94-109). Oxford University Press: New York.

[50] Gratch, J., Marsella, S., Wang, N., & Stankovic, B. (2009). Assessing the validity of appraisal-based models of emotion. In 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-8). IEEE.

[51] Yik, M., Russell, J. A., & Steiger, J. H. (2011).

[52] Lance, B., & Marsella, S. (2010). Glances, glares, and glowering: how should a virtual human express emotion through gaze? Autonomous Agents and Multi-Agent Systems, 20, 50-69.

---

[53] Yik, M., Russell, J. A., & Steiger, J. H. (2011).

[54] Zhang, S., Wu, Z., Meng, H. M., & Cai, L. (2010). Facial expression synthesis based on emotion dimensions for affective talking avatar. Modeling Machine Emotions for Realizing Intelligence: Foundations and Applications, 109-132.

[55] Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology, 76(5), 805–819.

[56] Izard, C. E. (1992), Basic emotions, relations among Emotions, and emotion-cognition relations, Psychological Review, Vol. 99, No. 3, 561-565.

[57] Quartz, S. R. (1999). The constructivist brain. Trends in cognitive sciences, 3(2), 48-57.

[58] Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. Current opinion in neurobiology, 23(3), 361-372.

[59] Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Social cognitive and affective neuroscience, 12(1), 1-23.

[60] Hudlicka, E. (2007). Reasons for emotions. Integrated models of cognition systems, 1, 263.

[61] Broekens, J., & DeGroot, D. (2004, November). Scalable and flexible appraisal models for virtual agents. In Proceedings of the International Conference on Computer Games, Artificial Intelligence, Design and Education (CGAIDE) (pp. 208-215).

[62] Dias, J., Mascarenhas, S., & Paiva, A. (2014). Fatima modular: Towards an agent architecture with a generic appraisal framework. Emotion modeling: Towards pragmatic computational models of affective processes, 44-56.

[63] Hudlicka, E. (2007). Reasons for emotions. Integrated models of cognition systems, 1, 263.

[64] Broekens, J., & DeGroot, D. (2004, November). Scalable and flexible appraisal models for virtual agents. In Proceedings of the International Conference on Computer Games, Artificial Intelligence, Design and Education (CGAIDE) (pp. 208-215).

[65] Broekens, J., & DeGroot, D. (2004, November).

[66] Broekens, J., & DeGroot, D. (2004, November).

[67] Broekens, J., & DeGroot, D. (2004, November).

---

[68] Sundström, P. (2005) Exploring the affective. University of Sweden, Stockholm, accessed on 1 October 2019 at http://www.diva-portal.org/smash/record.jsf?pid=diva2:1041047

[69] Meisner, S. (1987) Sanford Meisner on acting, Vintage Random House: New York, 9-10, 16-25, 26-38, 38-56, 57-77, 96-114, 136-147.

[70] Omicini, A., & Ossowski, S. (2003). Objective versus subjective coordination in the engineering of agent systems. In Intelligent Information Agents (pp. 179-202). Springer, Berlin, Heidelberg.

[71] Gerblick, J., (2020) Sony explains how The Last of Us 2 has such realistic facial animations. In Games Radar, August 28, 2020, 4 March 2023 at https://www.gamesradar.com/sony-explains-how-the-last-of-us-2-has-such- realistic-facial-animations/

[72] Boxer, S., (2013) How video games are transforming the film industry. In The Guardian, November 17, 2013, Accessed 4 March 2023 at https://www.theguardian.com/technology/shortcuts/2013/nov/17/video- games-transforming-film-industry

[73] Majek, D., (2021) The cinematisation of computer and console games. Bachelors Thesis, University of Stockholm, Autumn 2011, Accessed 4 March 2023, at https://www.diva portal.org/smash/get/diva2:756308/FULLTEXT01.pdf

[74] Paige, N., (2020) How to create smarter NPCs in games. in Medium, March 11, 2020, Accessed June 12. 2021, https://medium.com/@noahlandonpaige/how- to-create-smarter-npcs-in-games-10e384295f35

[75] Meisner, S. (1987) pp. 26-38.

[76] 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), pp. 428-433.

[77] Mac Namee, B., Cunningham, P. (2001) Proposal for an agent architecture for proactive persistent non player characters. In Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland: Trinity College, pp. 221-232.

[78] 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.

[79] Lim, M.Y., Dias, J., Aylett, R., Paiva, A., (2012) Creating adaptive affective autonomous NPCs, In Autonomous Agents and Multi-Agent Systems, New York, NY, USA: Springer, pp. 287–311.

[80] Ortony, A., Clore, G. L., and Collins, A., (1990). 34-58.

[81] Ochs, M., Sabouret, N., and Corruble, V. (2008) Modeling the dynamics of non- player characters' social relations in video games. In Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE), Menlo Park, CA, USA: AAAI Press, pp. 91-95

[82] Lisetti, C., and Hudlicka, E, (2015). 94- 109.

[83] Gratch, J., Marsella, S., Wang, N., and Stankovic, B. (2009). 1-8.

[84] Bakker, I., van der Voordt, T., Vink, P. and de Boon, J. (2014) Pleasure, arousal, dominance: Mehrabian and Russell revisited, In Current Psychology vol. 33, New York, NY, USA: Springer, pp. 405–421

[85] Scherer, K. R., (2005) What are emotions? And how can they be measured? In Social Science Information, vol. 44, no. 4, New York, NY, USA: Sage pp. 695-729

[86] Bentley, E, (1962) Who was Ribot? or: Did Stanislavsky know any psychology? The Tulane Drama Review, vol. 7, no. 2. Cambridge, UK: Cambridge University Press, pp. 127-129.

[87] Clurman, H. (1997) On Directing. New York: Fireside, pp.74-86.

[88] Meisner, S. (1987) pp. 26-38.

[89] Loyall, A. B., (1997) Believable agents: Building interactive personalities (No. CMU-CS-97-123), Carnegie-Mellon University, Pittsburgh, PA Department of Computer Science, accessed 30 June 2021 at: https://www.cs.cmu.edu/afs/cs/project/oz/web/papers/CMU-CS-97-123.pdf

[90] 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.

[91] Kapoor, A., (2015) Machine learning for affective computing: Challenges and opportunities, in The Oxford Handbook of Affective Computing, Oxford, UK: Oxford University Press, pp. 613-626.

[92] Khan, N. U., (2013) A comparative analysis of facial expression recognition techniques, In 3rd IEEE International Advance Computing Conference (IACC), New York, NY, US: IEEE, pp. 1262-1268.

[93] Panda, R., Zhang, J., Li, H., Lee, J., Lu, X., and Roy-Chowdhury, A. (2018). Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 579-595

[94] Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N. and Provost, E. (2016) MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. In IEEE Transactions on Affective Computing, vol. 8, no. 1, New York, NY, USA: IEEE, pp.67-80.

[95] Metallinou, A., Yang, Z., Lee, C.C., Busso, C., Carnicke, S. and Narayanan, S. (2016) The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations, In Language resources and evaluation, vol. 50, no. 3, New York, NY, USA: Springer, pp.497-521.

[96] Busso, C., Parthasarthy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. (2015) MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. In Transactions on Affective Computing, v. 10, no. 10, September 2015, New York, NY, USA: IEEE, pp. 1-16.

[97] Dupré, D., Krumhuber, E.G., Küster, D. and McKeown, G.J. (2020) A performance comparison of eight commercially available automatic classifiers for facial affect recognition, In Plos One, v. 15, no. 4, April 24, 2020, Accessed June 30, 2021, at https://doi.org/10.1371/journal.pone.0231968.

[98] Lewinski, P., den Uyl, T.M. and Butler, C. (2014) Automated facial coding: validation of basic emotions and FACS AUs in FaceReader, In Journal of Neuroscience, Psychology, and Economics, vol. 7, no. 4, Washington DC: APA, pp. 227-236.

[99] Kozasa, C, Fukutake, H., Notsu, H., Okada, Y., and Niijima, K. (2006). 428-433.

[100] Mascarenhas, S., Guimarães, M., Santos; P.A., Dias, J., Prada, R. and Paiva; A. (2021).

[101] Lim, M.Y., Dias, J., Aylett, R., Paiva, A., (2012).

[102] 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, pp. 619-623.

[103] Khorrami, P., Le Paine, T., Brady, K., Dagli, C. and Huang, T.S. (2016). 619-623.

[104] 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

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

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

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

[108] 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.

[109] 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. Bänziger, and E. B. Roesch Oxford Univ. Press, Oxford, England. 166–176.

[110] 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.

[111] 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.

[112] 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.

[113] 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

[114] Bänziger, T., Mortillaro, M., and Scherer, K.R., (2011).

[115] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S., (2008).

[116] 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.

[117] Meisner, S. (1987) pp. 26-38.

[118] 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.

[119] Barros, P., Churamani, N., Lakomkin, E., Siquiera, H., Sutherland, A. and Wermter. S. (2018).

[120] 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.

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

[122] Kozasa, C., Hiromiche. F., Notsu, H., Okada, Y., Niijima, K., (2006). 428-43.

[123] 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.

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

[125] 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.

[126] Bänziger, T.; Mortillaro, M.; and Scherer, K.R., (2011).

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

[128] 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.

[129] Meisner, S. (1987) pp. 26-38.

---

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

[131] Clurman, H. (1972) On directing. Fireside, New York, NY, USA. 80.

[132] Barros, P., Churamani, N., Lakomkin, E., Siquiera, H., Sutherland, A. and Wermter. S. (2018).

[133] Darwin, C., Prodger, P. (1872/1998). The expression of the emotions in man and animals. Oxford University Press, USA

[134] 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.

[135] Paier, W., Hilsmann, A., and Eisert, P. (2021). Example-based facial animation of virtual reality avatars using auto-regressive neural networks. IEEE Computer Graphics and Applications, 41(4), pp. 52-63.

[136] Paier, W., Hilsmann, A., and Eisert, P. (2020). Neural face models for example-based visual speech synthesis. In European Conference on Visual Media Production, pp. 1-10.

[137] Suwajanakorn, S., Seitz, S. M., Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics, 36(4), pp. 1-13.

[138] Wingenbach, T., Ashwin, C., Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial expression set – Bath Intensity Variation (ADFES-BIV), In PLoS ONE 11(1): e0147112.

[139] Skiendziel, T., Rösch, A. G., Schultheiss, O.C. (2019).

[140] 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.

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

[142] Darwin, C., and Prodger, P. (1872/1998). The expression of the emotions in man and animals. Oxford University Press, USA.

[143] 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

[144] 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.

[145] 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.

[146] Paier, W., Hilsmann, A., and Eisert, P. (2021).

[147] 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)

[148] Paier, W., Hilsmann, A., and Eisert, P. (2020).

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

[150] Krumhuber, E. G., Kappas, A., and Manstead, A. S. (2013).

[151] Sato, W., and Yoshikawa, S. (2004). Brief Report: The Dynamic aspects of emotional facial expressions. Cognition and emotion, 18(5).

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

[153] Bänziger, T.; Mortillaro, M.; and Scherer, K.R., (2011).

[154] 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

[155] 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.

[156] 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

[157] Metallinou, A.; Lee, C.; Busso, C.; Carnicke, S.; and Narayanan, S. (2010).

[158] 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.

[159] Barros, P.; Churamani, N., Lakomkin, E.; Siquiera, H., Sutherland, A.; and Wermter. S. (2018).

[160] Soleymani, M.; Larson, M.; Pun, T.; and Hanjalic, A. (2014) Corpus development for affective video indexing. In IEEE Transactions on Multimedia, 16(4).

[161] Cohn, J., Ambadar, Z., Ekman, P. (2007).

[162] Soleymani, M.; Larson, M.; Pun, T.; and Hanjalic, A. (2014).

[163] 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.

[164] 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).

[165] Loyall, A. B., (1997).

[166] 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.

[167] Ortony, A., Clore, G. L., and Collins, A. (1988). 34-58.

[168] Bates, J., Loyall, A. B., and Reilly, W., (1994).

[169] Kozasa, C., Hiromiche. F., Notsu, H., Okada, Y., Niijima, K., (2006). 428-43.

[170] Mascarenhas, S., Guimarães, M., Santos; P.A., Dias, J., Prada, R. and Paiva; A., (2021).

[171] Lim, M. Y., Dias, J., Aylett, R., & Paiva, A. (2012). Creating adaptive affective autonomous NPCs. Autonomous Agents and Multi-Agent Systems, 24, 287-311.

[172] 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.

[173] Lewinski, P., Den Uyl, T. M., and Butler, C. (2014). 227-236.

[174] 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.

[175] Stebbins, G. (1902). Delsarte system of expression. Edgar S. Werner.

[176] Hart, H. (2005). Do You See What I See? The Impact of Delsarte on Silent Film Acting. Mime Journal, 23(1), 184-199.

[177] O'Neill, R. M. (1927). The Science and art of speech & gesture: A comprehensive survey of the laws of gesture and expression, Founded on the art and life work of Delsarte, with his exercises.

[178] Whyman, R. (2022). The Science of acting in the Russian theatre at the beginning of the twentieth century–From the modern epoch to the avant-garde. Russian Literature.

[179] Marsella, S. C., Carnicke, S. M., Gratch, J., Okhmatovskaia, A., & Rizzo, A. (2006). An exploration of Delsarte's structural acting system. In Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6 (pp. 80-92). Springer Berlin Heidelberg.

[180] Nixon, M., Pasquier, P., & El-Nasr, M. S. (2010). DelsArtMap: Applying delsarte's aesthetic system to virtual agents. In Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10 (pp. 139-145). Springer Berlin Heidelberg.

[181] Neff, M. (2014). Lessons from the arts: what the performing arts literature can teach us about creating expressive character movement. Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters, 123-148.

[182] Leach, R. (2012). Meyerhold and biomechanics. In Twentieth-Century Actor Training (pp. 55-72). Routledge.

[183] Freud, S. (1921). A general introduction to psychoanalysis. Boni and Liveright.

[184] Bentley, E. (1962). Who was Ribot? Or did Stanislavsky know any psychology? Tulane Drama Review, 7(2), 127-129.

[185] Coger, L. I. (1964). Stanislavski changes his mind. Tulane Drama Review, 9(1), 63-68. https://doi.org/10.2307/1124778

[186] Stanislavski, C. (1989). An actor prepares. Routledge, 191.

[187] Hagen, U. (2008) Respect for acting. John Wiley & Sons.

[188]Arnold, M. B. (1960). Emotion and personality. Columbia University Press.

[189] Lazarus, R. S., & Folkman, S. (1984). Stress, appraisal, and coping. Springer publishing company.

[190] Arnold, M. B. (Ed.). (2013). Feelings and emotions: The Loyola symposium (Vol. 7). Academic Press.

[191] Ortony, A., Clore, G. L., and Collins, A. (1988).

[192] Ortony, A., Clore, G. L., and Collins, A. (1988).

[193] Bates, J., Loyall, A. B., and Reilly, W., (1994).

[194] Loyall, A. B., (1997).

[195] Reilly, W. S., & Bates, J. (1992). Building emotional agents. School of Computer Science, Carnegie Mellon University.

[196] Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2009, September). The OCC model revisited. In Proc. of the 4th Workshop on Emotion and Computing (p. 62). Palo Alto: Association for the Advancement of Artificial Intelligence.

[197] Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2009, September).

[198] Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2009, September).

[199] Strandberg-Long, P. (2019). The reaction in counter-action: how Meisner technique and active analysis complement each other. Stanislavski Studies, 7(1), 95-108.

[200] Dias, J., Mascarenhas, S., and Paiva, A., (2014).

[201] Dias, J., Mascarenhas, S., and Paiva, A., (2014).

[202] Darwin, C., & Prodger, P. (1998). The expression of the emotions in man and animals. Oxford University Press, USA.

[203] Deng, Weihong and Li, Shan (2023) Real-world Affective Faces Database , last accessed 21 February 2023, http://www.whdeng.cn/raf/model1.html

[204] Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971). Facial affect scoring technique: A first validity study.

---

[205] Ekman, P., & Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. Motivation and emotion, 10, 159-168.

[206] Ekman, P. (1992). 169-200.

[207] Ekman, P. (1992). 169-200.

[208] Kozasa, C., Hiromiche. F., Notsu, H., Okada, Y., Niijima, K., (2006). 428-43.

[209] Russell, J. A. (1995). Facial expressions of emotion: What lies beyond minimal universality? Psychological bulletin, 118(3), 379.

[210] Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology, 17(3), 715-734.

[211] Posner, J., Russell, J. A., & Peterson, B. S. (2005).

[212] Damasio, A. R. (1998). 83-86.

[213] Wingenbach, T.; Ashwin, C.; Brosnan, M. (2016)

[214] Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971).

[215]Mead, M. (1975). The appalling state of the human sciences: Review of Darwin and facial expression. Journal of Communication, 25, 210.

[216] Gross, J. J., & Feldman Barrett, L. (2011). Emotion generation and emotion regulation: One or two depends on your point of view. Emotion review, 3(1), 8-16.

[217] Barrett, L. F., Robin, L., Pietromonaco, P. R., & Eyssell, K. M. (1998). Are women the 'more emotional' sex? Evidence from emotional experiences in social context. Cognition & Emotion, 12(4), 555-578.

[218] Immordino-Yang, M. H., Yang, X. F., & Damasio, H. (2016). Cultural modes of expressing emotions influence how emotions are experienced. Emotion, 16(7), 1033.

[219] Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. Trends in cognitive sciences, 11(8), 327-332.

[220] Ekman, P. (1992). 169-200.

[221] Barrett, L. F., & Gross, J. J. (2001). Emotional intelligence: A process model of emotion representation and regulation.

[222] Ekman, P., and Cordaro, D. (2011). What is meant by calling emotions basic. Emotion review, 3(4), 364-370.

[223] Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A Survey on databases for multimodal emotion recognition and an introduction to the VIRI (Visible and InfraRed Image) Database. Multimodal Technologies and Interaction, 6(6), 47.

[224] Damasio, A. R. (1999). The feeling of what happens: Body and emotion in the making of consciousness. Houghton Mifflin Harcourt.

[225] Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. Trends in cognitive sciences, 11(8), 327-332.

[226] Immordino-Yang, M. H., Yang, X. F., & Damasio, H. (2016). 1033.

[227] Fan, Y., Lam, J. C., & Li, V. O. (2021). Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. Scientific reports, 11(1), 1-11.

[228] Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. Journal of personality and social psychology, 58(2), 342.

[229] Schiller, D. (2021). The face and the faceness: Iconicity in the early faciasemiotics of Paul Ekman, 1957–1978. Σημειωτκή-Sign Systems Studies, 49(3-4), 361-382.

[230] Smith, G. M. (2003).

[231] Smith, G. M. (2003).

[232] Berry, M., & Brown, S. (2022). The dynamic mask: Facial correlates of character portrayal in professional actors. Quarterly Journal of Experimental Psychology, 75(5), 936-953.

[233] Schiffer, S. (2023). Measuring emotion intensity: Evaluating resemblance in neural network facial animation controllers and facial emotion corpora. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2, pages 160-168.

[234] Schiffer, S. (2023). Measuring emotion velocity for resemblance in neural network facial animation controllers and their emotion corpora. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence - Volume 1, pages 240-248.

[235] Ekman, P., and Cordaro, D. (2011) 364-370.