Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-20)

Multimodal Player Affect Modeling with Auxiliary Classifier Generative Adversarial Networks

Nathan Henderson, Wookhee Min, Jonathan Rowe, James Lester
Department of Computer Science, North Carolina State University, Raleigh, NC 27545
{nlhender, wmin, jprowe, lester}@ncsu.edu

Abstract

Accurately detecting player affect is an important component of player modeling. Multimodal approaches to player modeling have shown significant promise because of their capacity to provide a multi-dimensional perspective on player behavior. However, obtaining sufficient data for training multimodal models of player affect presents significant challenges, including the prevalence of noisy, unbalanced, or missing data generated by multimodal sensor systems. To address this problem, we introduce a multimodal player affect modeling framework that improves player affect detection by using Auxiliary Classifier Generative Adversarial Networks (AC-GANs). We demonstrate the use of a Wasserstein distance-based approach for filtering synthesized data created in a data augmentation framework, and we investigate the effectiveness of the AC-GAN discriminator as an alternative approach for detecting player affect. Results show that AC-GAN based affective modeling outperforms baseline methods while enhancing player models through synthetic data generation and improved affect detection.

Introduction

Player modeling is critical for player-adaptive games that dynamically tailor gameplay to individual users (Yannakakis and Togelius 2018). Player modeling has been investigated across a range of applications, including interactive narrative (Wang et al. 2018) and procedural content generation (Summerville et al. 2018), as well as game balancing and difficulty adjustment (Zohaib 2018). An important aspect of player modeling is recognizing players' affective states during gameplay, such as engaged concentration, frustration, and boredom. Affect plays a key role in player engagement, and it provides a lens for understanding player experience and informing AI-enabled game designs that dynamically adjust to player emotions and subsequently create more engaging experiences.

Recent years have seen growing interest in multimodal approaches for detecting player affect. Multimodal affect detection has been shown to yield improved affect classifier performance and robustness compared to unimodal techniques (Bosch et al. 2016; Sun et al. 2014). As hardware sensors for tracking eye gaze (Min et al. 2017), body movement (Grafsgaard et al. 2012), and facial expression (Soleymani et al. 2015) have improved in price and accessibility, multimodal sensor systems have seen a growing role in computational models of player affect. In parallel, sensor-free affect detectors have emerged that eschew sensor data and instead rely upon interaction log data, including gameplay traces (Min et al. 2017) and keystroke data (Sun et al. 2014), to predict players' affective states. Interaction-based affect detection is particularly useful in contexts where physical sensors are too intrusive, too expensive, or are otherwise prohibitive. However, combining sensor-free and sensor-based approaches for multimodal affect detection has shown promise across a variety of domains and contexts (Bosch et al. 2016; Psaltis et al. 2018; Min et al. 2017).

Multimodal machine learning approaches to affect detection require large quantities of training data from each modality. However, multimodal sensor systems often suffer from several problems, including calibration issues, sensor noise, missing or imbalanced data, and data storage constraints. Similarly, interaction-based affect detectors must contend with issues such as hardware failure, software errors, and logging issues. These challenges can significantly impact the amount of data available to train multimodal affect detection models, raising concerns about overfitting and data sparsity that can adversely impact the accuracy and robustness of player affect models.

We address the issue of insufficient data for multimodal player affect detection with Auxiliary Classifier Generative Adversarial Networks (AC-GANs). AC-GANs, which can effectively model multimodal data distributions, are utilized to generate synthetic data consisting of posture and gameplay data captured from players interacting with a

serious game, *TC3Sim*, for training emergency medical skills. To ensure the quality of the augmented dataset produced by the AC-GAN, we demonstrate the effectiveness of a filtering method based on the Wasserstein distance metric to ensure the augmented multimodal data follows the original data sample's distribution (Vallender 1974). Finally, we demonstrate the effectiveness of using the AC-GAN discriminator network as a classification model for detecting players' run-time affective states during gameplay with *TC3Sim*.

Related Work

Recent years have seen growing interest in multimodal player modeling. Martinez et al. (2013) investigated the use of deep learning techniques for affect detection using several physiological modalities captured from users playing a prey/predator video game. Similar use of sensor-based and sensor-free modalities have informed the development of player engagement models across a range of games and contexts (Bosch et al. 2016; DeFalco et al. 2018; Psaltis et al. 2018). Related work has explored the relationship between physical and physiological behavioral patterns and self-reported gameplay experiences (Yannakakis and Hallam 2008). Multimodal player modeling has been investigated within game-based learning environments for tasks such as student goal recognition (Min et al. 2017), predicting problem-solving performance and gameplay outcomes (Liapis et al. 2019), and early prediction of engagement and cognitive load (Wiggins et al. 2018).

Models of player affect when engaging with games can be utilized to inform user-adaptive gameplay and enhance player experiences. DeFalco et al. (2018) explored the use of frustration detection models to provide adaptive feedback to players. Hernandez et al. (2014) used affective models to provide personalized game narratives based on the emotional progression of players. Affective models have also been used to create affect-responsive gameplay experiences that encourage positive player affect (Blom et al. 2014).

Generative models for data augmentation, such as generative adversarial networks (GANs) and conditional GANs, have seen increased usage within affective modeling (Zhu et al. 2018; Chatziagapi et al. 2019; Krokotsch and Böck 2019). Zhu et al. (2018) investigated the effectiveness of GAN-based data augmentation for emotion detection using convolutional neural networks. Krokotsch and Böck (2019) applied GANs within an unsupervised learning framework to generate synthetic affective speech data. Chatziagapi et al. (2019) used conditional GANs to resolve class imbalances in a similar speech-based emotion detection task. Qiu and Zhao (2018) applied AC-GANs as a denoising approach in a cognitive emotion recognition task. Although GANs have received significant attention in the affective computing community, the use of AC-GANs to augment multimodal data for player affect detection has not been explored.

The novelty of our work includes 1) the use of AC-GANs as a generative and discriminative approach to improve affect detection; 2) the use of a Wasserstein distance-based filtering mechanism to guide generation of synthetic affect data; 3) the application of AC-GAN based affective modeling to posture and gameplay interaction data (in contrast to facial expression or speech data); and 4) our focus on multimodal data augmentation to model player emotions arising naturalistically during interactions with a serious game.

Dataset

The dataset we use to investigate GAN-based multimodal affect detection consists of data from students engaged with a serious game for emergency medical skills training called *TC3Sim*. The data was collected during a study conducted at the United States Military Academy with 119 first-year cadets (83% male, 17% female). While the cadets engaged with *TC3Sim*, two researchers observed the participants and recorded observations of students' affective states in accordance with the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) (Ocumpaugh et al. 2015). These observations serve as target labels for affect recognition. A Microsoft Kinect for Windows (version 1) was positioned in front of each player to capture skeletal vertex coordinate data on the player's posture and upper-body movement during gameplay. In addition, time-stamped interaction logs from player actions in *TC3Sim* were recorded. We utilize this dataset to investigate GANs for generating synthetic data, as well as to train and validate multimodal affect detection models for player experiences with *TC3Sim*.

TC3Sim Game-Based Learning Environment

TC3Sim is a widely used serious game for training tactical combat casualty care skills (Figure 1). During the game, players assume the role of a combat medic in a simulated military training scenario. Players navigate a series of simulated scenarios centered upon administering emergency medical care to injured computer-controlled teammates in accordance with U.S. Army medical procedures. Each player completed four training scenarios, ranging from simple application of a tourniquet to a scenario where the injured character expires regardless of the medical care administered by the player. Players engaged with *TC3Sim* individually, and sessions lasted approximately one hour.

Affect Observation Protocol

A pair of trained observers followed the BROMP observation protocol (Ocumpaugh et al. 2015) to label players' affective states during interactions with *TC3Sim*. The two

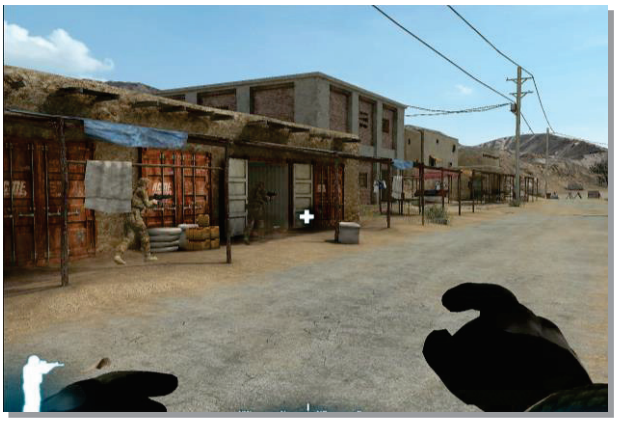

Figure 1: *TC3Sim* Serious Game for Emergency Medical Skills Training.

observers achieved an inter-rater agreement exceeding 0.6 in terms of Cohen's kappa. During the study, six distinct affective states were recorded: boredom, confusion, engagement, frustration, surprise, and anxiety. After removing any affective observations recorded during times when participants were not using *TC3Sim* and observations in which the two coders disagreed with one another, there were 755 observations remaining. 435 of the BROMP observations were labeled as *engaged* (M = .58, SD = .24), 174 as *confused* (M = .23, SD = .19), 73 as *bored* (M = .10, SD = .16), 32 as *frustrated* (M = .04, SD = .18), 29 as *surprised* (M = .04, SD = .05), and 12 as anxious (M = .02, SD = .09). Due to the low number of observed instances of anxious, we exclude this affective state from our work.

Methodology

As a response to the relatively small size of the dataset, we seek to improve the accuracy of multimodal affect detection models by increasing the amount of available training data through synthetic data generation. Because of significant class imbalance issues in the data, we generate augmented data based on the original data points and only add generated instances of minority class instances to the dataset. After the dataset augmentation process, we train five binary classifiers for the five target affective states: *bored*, *confused*, *engaged concentration*, *frustrated*, and *surprised*.

Feature Representation

The Kinect captured 3D coordinate data for 91 vertices. As the basis for the feature engineering process, we selected three vertices to track student posture: *top_skull*, *center_shoulder*, and *head*. The selection of these particular vertices is supported by prior work on Kinect-based affective modeling (Grafsgaard et al. 2012). From this data, we distilled 73 posture-based features providing a summative perspective on participants' body posture. For each of the three vertices, 18 statistical features were generated. The features included the most recent observed distance, most recent Z-coordinate, minimum observed distance, maximum observed distance, median distance, and variance in observed distances. Distance was defined as the Euclidean distance of the vertex from the Kinect. In addition, summative features for each vertex were calculated using the minimum distance, maximum distance, median distance, and variance in distance observed across the preceding 5, 10, and 20 second windows prior to each affect observation. In addition to these 54 features, we generated several *net_change* features representing the total change in position and distance from the Kinect sensor over time windows of 3 and 20 seconds. Finally, three features indicating whether the student was sitting forward, upright, or backwards were computed using the median distance of each *head* vertex for each workstation and the current distance of the *head* vertex from the Kinect sensor. These features were then computed across time windows of 5, 10, and 20 seconds, in addition to the entire gameplay session up to the current BROMP observation.

Additionally, we generated 48 temporal features based on the "velocity" of the *head* vertex. These features were generated by calculating the delta values between two consecutive Kinect readings and then calculating the mean, median, maximum, and variance of the corresponding velocity values across time windows of 3, 5, 10, and 20 seconds prior to each BROMP observation. This feature engineering was only completed for a single vertex (*head*) due to the large number of features generated for a single vertex.

Interaction-based features were generated from gameplay data logs collected during player interactions with *TC3Sim*. These features capture player actions performed during the gameplay sessions as well as information about the status of non-player characters (NPC) throughout the course of the game. Features extracted from NPC-based patients include changes in systolic blood pressure, exposed wounds, lung volume, remaining blood volume, and bleed rate. Features based on player gameplay actions include checking a patient's vital signs, conducting a blood sweep, corresponding with a patient, or requesting a medical evacuation. Each feature was cumulatively calculated over the preceding 20 seconds prior to each BROMP observation. This process produced 39 distinct interaction-based features. Additional analysis of the predictive value of specific features in this dataset can be found in (Henderson et al. 2020).

Auxiliary Classifier Generative Adversarial Networks

Auxiliary Classifier Generative Adversarial Networks (AC-GANs) (Odena, Olah, and Shlens 2017) are an extension of generative adversarial networks which consist of

two deep neural network models, a *generator* and a *discriminator*, that compete against one other in an adversarial fashion within a zero-sum setting to generate synthetic data that resembles the original data used for training (Goodfellow et al. 2014). Using a random Gaussian noise vector as input, the generator aims to synthesize realistic data that deceives the discriminator, whose task is to accurately distinguish between "real" data from the training data and "fake" data produced by the generator. The discriminator loss is backpropagated through both components of the GAN, with theoretical convergence achieved when the components' losses reach a Nash equilibrium. Conditional GANs extend this model by providing associated information to both the generator and discriminator, such as a class label associated with the desired synthetic output (Mirza and Osindero 2014). AC-GANs deviate from the conditional GAN architecture by allowing the discriminator to predict the class label of the generated sample as well as the data source (i.e., "real" or "fake" status). The generator aims to minimize the ability of the discriminator to distinguish between real and fake data, while maximizing its ability to predict the class label of the generated data. This often leads to a more stabilized training process and also allows the GAN to learn a latent space representation that does not rely on the class label as input, unlike a standard conditional GAN. The discriminator's ability to be trained to predict the class label of the generated data lends itself for additional use as an affect classification model, a property that is investigated within this work.

Wasserstein Filtering

To ensure that our synthetic data accurately reflects the distribution of the original dataset, we utilize a filtering process based on the Wasserstein metric (Vallender 1974). Also known as the "earth mover's distance," this metric is grounded in optimal transport theory and is a method to quantify the distance between two probability distributions. We select this metric due to its ability to account for both the probability density of the synthetic data compared to the original distribution and also the distance within a defined metric space, giving it an advantage over related methods such as Kullback-Leibler (KL) divergence.

Once the AC-GAN is used to generate batches of 50 augmented data samples, using a Gaussian noise vector and the minority class label as the conditioning variable to the generator, the average Wasserstein distance between each feature and the corresponding feature in the original dataset is computed across all features for each generated sample in a single batch. After this process is repeated for 10 batches, the batch with the lowest average Wasserstein distance is selected for inclusion in the augmented dataset. This process continues iteratively until all classes in the dataset are uniformly distributed. This method ensures that the synthetic data contains an appropriate variance level beneficial for the affective models while not creating closely identical examples of the original minority data, which could induce overfitting in the affective models.

Affect Model Evaluation

Using the balanced datasets, we evaluate several different machine learning models and determine which model produces binary affect classifiers with the greatest accuracy. For each affective state (e.g., engaged, confused, bored, frustrated, surprised), a "raw" dataset is constructed with binary labels indicating whether or not a given BROMP observation is a positive instance of that affective state. We use five different models for each affective state: support vector machine (SVM), random forest (RF), Gaussian naïve Bayes (NB), logistic regression (LR), and multilayer perceptron (MLP). For evaluations involving a GAN model, the trained discriminator was retained from the data augmentation phase and trained further using the same data as the other five models. The discriminator's predictions of the class label were used during the affect model evaluations. The models' classification performance is measured in terms of Area Under the Curve (AUC) as the primary evaluation metric to account for model correctness in the face of class imbalance. We also include the predictive accuracy as well as the F1 score, recall, and precision for each model to illustrate the tradeoffs inherent among different evaluation metrics. To serve as a baseline against which we compare our affective models' classification performance, we train the same set of models on the raw normalized dataset without any prior data augmentation.

We compare our data augmentation framework against two common approaches for resolving class imbalance issues in affective modeling: minority cloning and Synthetic Minority Over-Sampling Technique (SMOTE). Minority cloning involves the duplication of each instance of the minority class at a rate that brings the classes closer to a balanced distribution. SMOTE selects a minority data sample at random and then linearly interpolates synthetic data points between the selected point and another randomly selected minority sample chosen by a K-nearest neighbor algorithm (Chawla et al. 2002).

The models were trained using 5-fold cross-validation with data splits maintained on a player level to prevent data leakage among individual gameplay sessions. 5-fold cross-validation was chosen to maintain an adequate number of positive instances of each affective state within each validation fold. The class distribution within each fold was maintained using stratified sampling. Prior to training, the training dataset was normalized so that each feature's range fell between [-1, 1]. Following normalization, feature selection was performed by retaining a subset of features that contained the highest chi-squared test values with the class variable. Normalization, feature selection, and class balancing

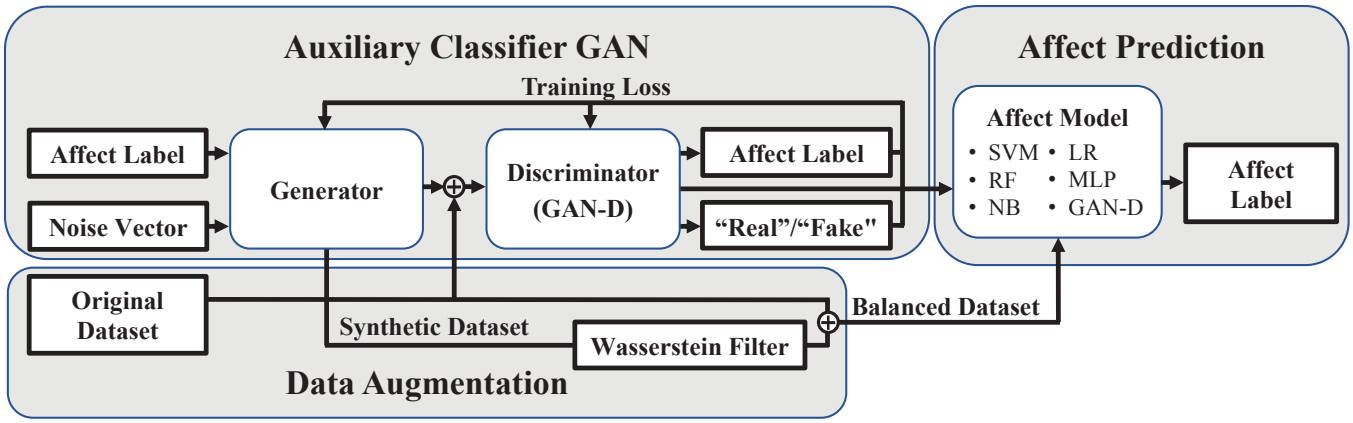

Figure 2: Data Augmentation for Affect Model Training Using an AC-GAN.

took place with the training set of each cross-validation step to ensure that data leakage did not occur during baseline or GAN upsampling. The data augmentation process using the AC-GAN is illustrated in Figure 2.

Results

For each upsampling technique and affective state, we evaluate five classification techniques (SVM, RF, NB, LR, and MLP) in addition to the AC-GAN discriminator. The highest-performing models are presented in Table 1, with the results of the optimal classifiers in terms of area under the curve (AUC) shown in bold. We compare our results to the baseline classifiers trained on the original dataset as well as classifiers trained on augmented data that underwent different forms of upsampling.

AC-GAN data augmentation outperformed each of the other upsampling approaches in terms of AUC with the exception of *bored*. For each of the 4 remaining affective states, the AC-GAN combined with the Wasserstein filtering (AC-GAN-W) outperformed the standard AC-GAN in 3 of 4 cases. Additionally, the AC-GAN discriminator was selected as the optimal affect model in 5 of 10 possible cases across the two AC-GAN upsampling tests for the 5 distinct affective states (an AC-GAN discriminator was not trained for the baseline, cloning, or SMOTE upsampling experiments). This 50% selection rate was the highest selection rate among all affective models, compared to 20% (5/25) for SVM, 32% (8/25) for naïve Bayes, 12% (3/25) for logistic regression, and 16% (4/25) for multilayer perceptron.

Discussion

The results indicate that AC-GANS are the highest-performing upsampling approach for 4 of the 5 affective states. The lone exception, *bored*, demonstrated very high AUC values for all of the upsampling techniques. In this case, the AC-GAN upsampling technique outperformed the baseline, but the other upsampling techniques induced higher AUC scores from the models than the AC-GANs. This may be due to predictive anomalies in the data or boredom-specific behavioral cues, which warrant further investigation.

The impact of the AC-GAN augmentation was more significant for *frustrated* and *surprised*, the two most imbalanced classes in the dataset. One explanation for this behavior is that the data points belonging to the minority class are likely highly localized, as these instances of *frustrated* and *surprised* comprise 4.2% and 3.8% of the total dataset, respectively. Because SMOTE is based on a K-nearest neighbor approach, the augmented data will be contained within the same range as the original data points.

Minority cloning does not introduce variance during data augmentation, and as a consequence, may cause classifiers to overfit minority data, which is likely to harm affect detector accuracy. While SMOTE introduces some variance during data synthesis, it is limited by its dependence upon producing samples using linear interpolation. This could also lead to the predictive model conforming to the localization of the minority class and overfitting of the model. One aspect of using generative models such as AC-GANs for data augmentation is their capacity to model complex relationships between various data attributes through non-linear transformations and generate synthetic data according to the underlying distributions while still maintaining a beneficial amount of variance for the classifiers. It should be noted that the results indicate that the Wasserstein distance-filtering approach is an effective method of enforcing an adequate variance level in the synthetic data while still retaining accurate modeling of the original data distributions. This allows the AC-GAN-based data augmentation process to be more robust when encountering heavily skewed data.

Of note is the performance of the AC-GAN discriminator as the most frequently selected optimal affect model. Using subsets of real and artificial data from the generator, the weights of the AC-GAN are initially trained within a multi-

Table 1: Optimal Models for Each Combination of Modalities and Affective States.

	Bored					Confused					
Upsampling	**Model**	**AUC**	**Acc.**	**F1**	**Prec.**	**Rec.**	**Model**	**AUC**	**Acc.**	**F1**	**Prec.**	**Rec.**
Baseline	NB	0.701	0.542	0.277	0.176	0.889	NB	0.520	0.509	0.229	0.146	0.539
Cloning	SVM	0.815	0.827	0.461	0.354	0.794	NB	0.518	0.507	0.229	0.145	0.539
SMOTE	**LR**	**0.817**	**0.839**	**0.474**	**0.371**	**0.785**	NB	0.518	0.513	0.228	0.146	0.527
AC-GAN	NB	0.774	0.831	0.421	0.324	0.696	MLP	0.533	0.662	0.232	0.257	0.296
AC-GAN-W	GAN-D	0.662	0.802	0.330	0.272	0.494	**GAN-D**	**0.543**	**0.588**	**0.326**	**0.297**	**0.454**
	Engaged Concentration						**Frustrated**					
Upsampling	**Model**	**AUC**	**Acc.**	**F1**	**Prec.**	**Rec.**	**Model**	**AUC**	**Acc.**	**F1**	**Prec.**	**Rec.**
Baseline	SVM	0.569	0.610	0.705	0.621	0.821	MLP	0.573	0.920	0.132	0.125	0.195
Cloning	SVM	0.569	0.610	0.705	0.621	0.821	LR	0.614	0.717	0.119	0.070	0.500
SMOTE	SVM	0.554	0.578	0.650	0.613	0.707	LR	0.654	0.847	0.194	0.128	0.445
AC-GAN	**GAN-D**	**0.575**	**0.616**	**0.720**	**0.622**	**0.862**	GAN-D	0.699	0.818	0.183	0.121	0.565
AC-GAN-W	GAN-D	0.571	0.523	0.461	0.557	0.432	**MLP**	**0.748**	**0.664**	**0.173**	**0.097**	**0.840**

	Surprised					
Upsampling	**Model**	**AUC**	**Acc.**	**F1**	**Prec.**	**Rec.**
Baseline	NB	0.517	0.389	0.077	0.041	0.656
Cloning	NB	0.530	0.388	0.080	0.043	0.685
SMOTE	SVM	0.501	0.801	0.054	0.033	0.176
AC-GAN	NB	0.562	0.679	0.077	0.044	0.436
AC-GAN-W	**MLP**	**0.617**	**0.578**	**0.105**	**0.062**	**0.657**

task framing, meaning that the discriminator learns to not only distinguish trends of the real and artificial data, but also the binary class label of each sample. This factor may play a role in the enhanced performance of the AC-GAN discriminator as an affect model.

Conclusion

Detecting player affect is a key component of player modeling. Computational models of affect enable the creation of affect-sensitive games that dynamically adapt to player emotions. Affective modeling has significant implications for enabling dynamic difficulty adjustment based on player emotions, virtual agents that dynamically respond to players' emotions at run-time, and player analytics that can be utilized to inform game design decisions There is growing evidence suggesting that multimodal affect detection is an effective approach for modeling player affect in digital games. However, obtaining sufficient data to train multimodal machine learning-based models of player affect detection is challenging, as it is often hindered by issues such as sensor noise, calibration issues, and mistracking.

We have presented a novel approach to multimodal affect detection that leverages data augmentation through the use of auxiliary classifier generative adversarial networks (AC-GANs) and utilizes Wasserstein distance as a metric to filter generated data. We demonstrate the effectiveness of this approach with a serious game for emergency medical skills training called *TC3Sim*. We also demonstrate the effectiveness of using the AC-GAN discriminator as a higher-performing alternative to standard affect detector modeling techniques. Results of the evaluation have shown that our method induces higher predictive performance than two common class imbalance resolution methods including SMOTE and minority cloning on four of five targeted affective states with respect to AUC.

The results suggest several possible avenues for future work. Additional generative modeling techniques merit further exploration for data augmentation, including different discriminator and generator architectures for AC-GANs, different GAN architectures such as Wasserstein Generative Adversarial Networks (W-GANs), and other generative modeling techniques such as variational autoencoders. Another promising direction is to investigate additional data filtration methods and modeling techniques for affect recognition. Finally, it will be important to investigate the run-time integration of multimodal affective models into player-adaptive games designed to enrich player experiences through affect-sensitive interventions.

Acknowledgements

The research was supported by the U.S. Army Research Laboratory under cooperative agreement #W911NF-13-2-0008. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army.

References

Blom, P.; Bakkes, S.; Tan, C.; Whiteson, S.; Roijers, D.; Valenti, R.; and Gevers, T. 2014. Towards Personalised Gaming via Facial Expression Recognition. In *Proceedings of the Tenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 30–36. Palo Alto, CA.: AAAI Press.

Bosch, N.; D'Mello, S.; Baker, R.; Shute, V., Ventura, M.; Wang, L.; and Zhao, W. 2016. Detecting Student Emotions in Computer-Enabled Classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 4125–4129.

Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; and Narayanan, S. 2019. Data Augmentation using GANs for Speech Emotion Recognition. *Proc. Interspeech 2019*, 171-175.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16(1): 321–357.

DeFalco, J.; Rowe, J.; Paquette, L.; Georgoulas-Sherry, V.; Brawner, K.; Mott, B.; Baker, R.; and Lester, J. 2018. Detecting and Addressing Frustration in a Serious Game for Military Training. *International Journal of Artificial Intelligence in Education* 28(2): 152–193.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2672-2680.

Grafsgaard, J.; Boyer, K.; Wiebe, E; and Lester, J. 2012. Analyzing Posture and Affect in Task-Oriented Tutoring. In *Proceedings of the International Conference of the Florida Artificial Intelligence Research Society*, 438–443.

Henderson, N.; Rowe, J.; Paquette, L.; Baker, R.; and Lester, J. 2020. Improving Affect Detection in Game-Based Learning with Multimodal Data Fusion. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence in Education*, 228-239.

Hernandez, S.; Bulitko, V.; and Hilaire, E. 2014. Emotion Based Interactive Storytelling with Artificial Intelligence. In *Proceedings of the Tenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 146–152. Palo Alto, CA.: AAAI Press.

Krokotsch, T., and Böck, R. 2019. Generative Adversarial Networks and Simulated+ Unsupervised Learning in Affect Recognition from Speech. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction,* 28-34.

Liapis, A.; Karavolos, D.; Makantasis, K.; Sfikas, K.; and Yannakakis, G. 2019. Fusing Level and Ruleset Features for Multimodal Learning of Gameplay Outcomes. In *Proceedings of the 2019 IEEE Conference on Games,* 1-8. Piscataway, NJ.: IEEE.

Martinez, H.; Bengio, Y.; and Yannakakis, G. 2013. Learning deep Physiological Models of Affect. *IEEE Computational Intelligence Magazine* 8(2): 20-33.

Min, W.; Mott, B.; Rowe, J.; Taylor, R.; Wiebe, E.; Boyer, K.; and Lester, J. 2017. Multimodal Goal Recognition in Open-World Digital Games. In *Proceedings of the Thirteenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 80-86. Palo Alto, CA.: AAAI Press.

Mirza, Mehdi; and Osindero, S. 2014. Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784. Ithaca, NY: Cornell University Library.

Ocumpaugh, J.; Baker, R.; and Rodrigo, M. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional Image Synthesis with Auxiliary Classifier GANS. In *Proceedings of the 34th International Conference on Machine Learning*, 2642-2651.

Psaltis, A.; Apostolakis, K.; Dimitropoulos, K.; and Daras, P. 2018. Multimodal Student Engagement Recognition in Prosocial Games. *IEEE Transactions on Games* 10(3): 292-303.

Qiu, J., and Zhao, W. 2018. Data Encoding Visualization Based Cognitive Emotion Recognition with AC-GAN Applied for Denoising. In *Proceedings of the 17th International Conference on Cognitive Informatics and Cognitive Computing*, 222-227.

Soleymani, M.; Asghari-Esfeden, S.; Fu, Y.; and Pantic, M. 2015. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing* 7(1): 17–28.

Summerville, A.; Snodgrass, S.; Guzdial, M.; Holmgård, C.; Hoover, A.; Isaksen, A.; Nealen, A.; and Togelius, J. 2018. Procedural Content Generation via Machine Learning. *IEEE Transactions on Games* 10(3): 257-270.

Sun, H.; Huang, M.; Ngai, G.; and Chan, S. 2014. Nonintrusive Multimodal Attention Detection. In *Proceedings of the 7th International Conference on Advances in Computer-Human Interactions*, 192–199.

Vallender, S. 1974. Calculation of the Wasserstein Distance between Probability Distributions on the Line. *Theory of Probability & Its* Applications 18(4): 784-786.

Wang, P.; Rowe, J.; Min, W.; Mott, B.; and Lester, J. 2018. High-Fidelity Simulated Players for Interactive Narrative Planning. In *Proceedings of the Twenty-Seventh Joint Conference on Artificial Intelligence,* 3884-3890.

Wiggins, J.; Kulkarni, M.; Min, W.; Mott, B.; Boyer, K.; Wiebe, E.; and Lester, J. 2018. Affect-Based Early Prediction of Player Mental Demand and Engagement for Educational Games. In *Proceedings of the Fourteenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 243-249. Palo Alto, CA.: AAAI Press.

Yannakakis, G.; and Hallam, J. 2008. Entertainment Modeling through Physiology in Physical Play. *International Journal of Human-Computer Studies* 66(10): 741-755.

Yannakakis, G., and Togelius, J. 2018. Modeling Players. In *Artificial Intelligence and Games*, 203-255. New York, NY.: Springer.

Zhu, X.; Liu, Y.; Li, J.; Wan, T.; and Qin, Z. 2018. Emotion Classification with Data Augmentation using Generative Adversarial Networks. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 349-360. New York, NY: Springer.

Zohaib, M. 2018. Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review. *Advances in Human-Computer Interaction* (2018).