Recognizing Emotional Expression in Game Streams

Shaghayegh Roohi
Aalto University
Espoo, Finland
shaghayegh.roohi@aalto.fi

Elisa D. Mekler
Aalto University
Espoo, Finland
elisa.mekler@aalto.fi

Mikke Tavast
Aalto University
Espoo, Finland
mikke.tavast@aalto.fi

Tatu Blomqvist
Aalto University
Espoo, Finland
tatu.blomqvist@aalto.fi

Perttu Hämäläinen
Aalto University
Espoo, Finland
perttu.hamalainen@aalto.fi

ABSTRACT

Gameplay is often an emotionally charged activity, in particular when streaming in front of a live audience. From a games user research perspective, it would be beneficial to automatically detect and recognize players' and streamers' emotional expression, as this data can be used for identifying gameplay highlights, computing emotion metrics or to select parts of the videos for further analysis, e.g., through assisted recall. We contribute the first automatic game stream emotion annotation system that combines neural network analysis of facial expressions, video transcript sentiment, voice emotion, and low-level audio features (pitch, loudness). Using human-annotated emotional expression data as the ground truth, we reach accuracies of up to 70.7%, on par with the inter-rater agreement of the human annotators. In detecting the 5 most intense events of each video, we reach a higher accuracy of 80.4%. Our system is particularly accurate in detecting clearly positive emotions like amusement and excitement, but more limited with subtle emotions like puzzlement.

CCS Concepts

•Human-centered computing → **HCI design and evaluation methods;** **•Computing methodologies** → *Machine learning;*

Author Keywords

Facial Expression; Games; Player Experience; Emotion; Neural Network; Sentiment Analysis

INTRODUCTION

Modern game development makes extensive use of large-scale online game testing through services like PlayTestCloud and Usertesting.com. This poses a need for better tools for summarizing and exploring the large quantities of playtest videos. A related and yet underexplored data trove is provided by game streaming videos publicly available through services like YouTube and Twitch. Automatic analysis of such video data, e.g., recognizing player emotions based on facial expressions and voice, has potential in providing valuable insights for both game development and player experience research (e.g., [25, 27, 29]).

To advance automated analysis of player emotion from video, we investigate the following research question: *Can one replicate how humans annotate game stream videos, when they are asked to detect and label moments with emotion, and identify the most emotional events of each video?* If human annotations can be predicted by an automated system, such a system should have applications in stream and gameplay highlight detection and in player experience research, e.g., in selecting gameplay videos or video segments for closer inspection after a playtest session. We assume that similar analysis methods may work for both streams and playtest videos, especially if playtesters are asked to think aloud or otherwise narrate their experience.

Our contribution is twofold:

- We present a new dataset of human-annotated emotional expression in game streams, including a total of 17 videos, 11 hours, and 2015 emotional events. Each video was annotated by two different persons to allow assessing the consistency of human annotations. On average, the dataset has approximately one annotated event for each 40 seconds of video.
- We propose a novel neural network system trained to mimic the human annotation behavior, illustrated in Figure 1. Our system is multimodal, combining four input types: facial expressions, video transcript sentiment analysis, voice emotion analysis, and additional voice features like loudness and pitch. Previous automatic gameplay emotion analysis approaches primarily focus on a single modality like facial expressions and compute emotion changes at predetermined events [29] or implement highlight detection with hand-tuned parameters without a dataset for assessing model accuracy [27].

CHI PLAY '19, October 22–25, 2019, Barcelona, Spain.
Copyright © 2019 Association of Computing Machinery.
ACM ISBN 978-1-4503-6688-5/19/10 ...$15.00.
http://dx.doi.org/10.1145/3311350.3347197

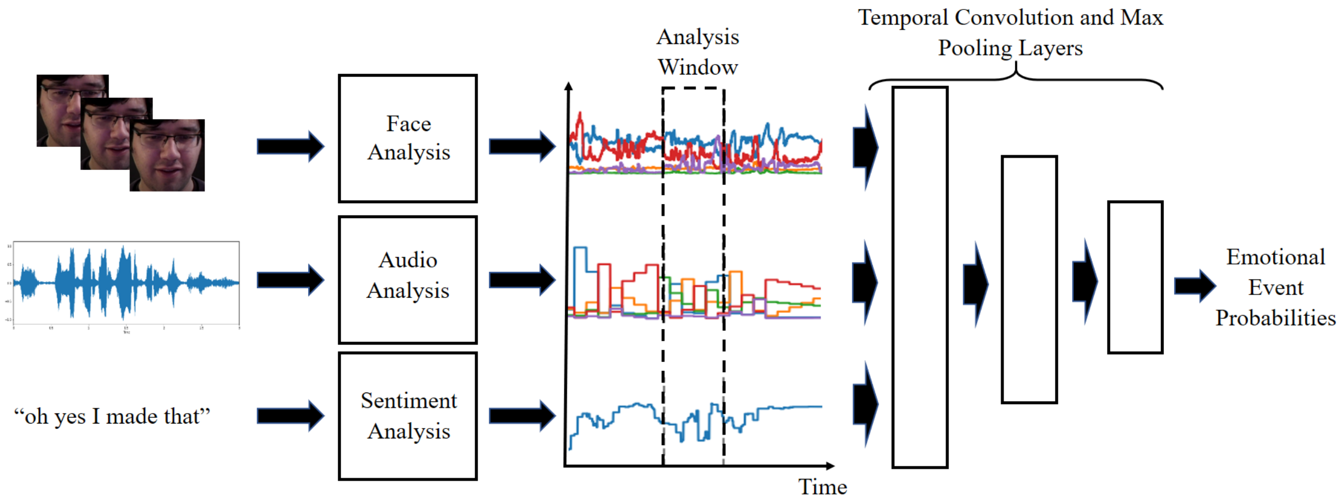

Figure 1. Our analysis pipeline. The video, audio, and sentiment analysis modules output time series signals such as subtitle sentiment positivity and emotion class probabilities for facial expressions. These are collected together from each analysis window and fed through a temporal convolutional neural network trained to output the probabilities of the emotional event labels of our dataset, also including "no event".

Our evaluation indicates that the proposed system is well suited for detecting emotional moments, in particular the most intense ones, but more work is needed in more fine-grained emotion analysis. On the other hand, the annotation task appears likewise difficult for humans. Based on inter-rater agreement scores, our annotators are fairly consistent in what events or moments they annotate, but disagreement grows with more fine-grained emotion labeling.

In the most simple "no event" vs. "emotional event" detection case, our system achieves a validation accuracy of 70.7%, on par with the inter-rater agreement of the human annotators.

The dataset is provided as the paper's supplementary material. While the game streams themselves cannot be included in the dataset because of copyright reasons, we provide timed YouTube links for each annotated event.

RELATED WORK

Yannakakis and Paiva [37] argued that "one cannot dissociate games from emotions" (p. 459). Indeed, many studies equate (positive) emotions with positive player experience ([22], e.g., [35]) and consider them key contributors to fun [18].

A variety of methods have therefore been suggested for analyzing player emotions. For example, in her observational studies on what makes games fun, Lazzaro [18] applied Ekman and Rosenberg's Facial Action Coding System (FACS) [13] to link different player experiences to specific emotional expressions. The experience of hard fun, for example, is characterized by players expressing "fiero", i.e., triumph at mastering a challenge [12], where players smile and lift their arms and body upwards. However, given the time and resources necessary to learn and manually apply FACS, the method is not feasible for evaluating hours of player-game interaction.

Several researchers also demonstrated the utility of biometrics for assessing the experience of players [9, 20, 23, 35] and streamers [28]. A major advantage of biometrics being that they allow for the continuous, real-time assessment of players' emotional reactions during gameplay. Of particular interest in the context of our work is the use of facial electromyography (EMG), which measures activation of the different facial muscles to assess emotional valence (i.e., how pleasant or unpleasant we experience a given event). Van den Hoogen et al. [35], for instance, combined EMG during play with a self-report questionnaire, where upon watching a replay participants had to rate their experience of valence. The study found that although participants self-reported dying in-game to be a negative event, they expressed positively valenced emotional reactions while playing. In another study, Mirza-Babaei et al. [23] introduce the Biometric Storyboard, a games user research tool that combines EMG with video observation and gamelogs. This approach visualizes players' emotional reactions (as measured by biometrics) over pre-defined key game events (i.e., events that the game designers are particularly interested in), which can then be shown playtesters to elicit additional feedback. Similarly, Tan et al. [34] combined different physiological measures (incl. EMG) with video-cued retrospective think-aloud to analyze the player experience of different pre-selected game events (e.g., approaching a lift in *Portal*). Notably, they found that think-aloud could detect noteworthy game events that were not discernible from the physiological measures. Conversely, biometrics could detect notable gameplay events, which players did not comment on and would have gone unnoticed otherwise. For instance, EMG revealed instances where players experienced positively valenced anticipation as they explored the game level. Taken together, the aforementioned research suggests that recording players' emotional expressions can support retrospective think-aloud procedures through assisted recall by presenting players with the most emotionally salient events.

However, while the aforementioned studies showcase the benefits of analyzing players' emotional expressions through biometrics, physiological measures can be disruptive to the player experience [34], require considerable equipment and training [9, 34], and remain difficult to reliably employ with large-scale sample sizes, such as for online playtesting. Consequently, a growing body of research has emerged around automated analysis of emotional expression based on a player's facial expressions and voice captured on video. Automatic emotion detection has been argued to be useful for analyzing the PX without interrupting the play session [37], to assess whether the intended emotional experience has been achieved and identify areas for improvement, as well as create emotionally-adaptive games to personalize the PX [37].

Automated Emotion Analysis from Video and Audio

In this paper, we aim for an automatic system for emotion detecting and classification suitable for large-scale online data collection. Considering the input devices prevalent currently and in the near future, the primary data sources of interest are player facial video and speech audio.

With deep convolutional neural networks, analysing player facial expressions from video is relatively robust and straightforward. Roohi et al. [29] showed that such visual analysis can produce similar findings as facial EMG [35], e.g., that player dying can elicit strong positive/happy expressions. Their network was trained on a publicly available dataset of 50k faces labeled by human annotators according to the 6 basic emotions and the associated facial expressions of Ekman [11, 12]. Murray et al. [25] applied facial expression analysis to key content events in a story-driven game, measuring engagement and valence.

In the work above, knowledge of game events helps in emotion analysis, providing anchor points around which one may hypothesize that players experience changes in their emotions. However, such event data may not be available, for example, when studying data collected from a third-party game with no instrumentation code. Although Intharah et al. [16] have shown that game events can be recognized from gameplay video using deep neural networks, such networks trained on one game do not necessarily generalize to other games.

An alternative research approach to investigating the emotions elicited by game events is attempting to predict post-game player experience measures from in-game behavior. Tan et al. [33] showed a relation between facial expressions and Game Experience Questionnaire dimensions. Shaker and Shaker [30] have also shown that post-game engagement measures can be predicted from players' in-game non-verbal behavior, especially when analysing behavior near the end of the game.

This paper utilizes video game streams as the data source. Streams are relevant to game emotion research because streamers are often emotionally expressive and even narrate their emotions aloud to the spectators. In contrast, the data of Roohi et al. [29] shows that for most game events, non-streamer players may not express any emotion, the expressionlessness possibly due to high concentration. Recently, Ringer et al. [27] demonstrated the potential of combined facial expression and audio analysis of game streamers in automated highlight recognition. In this paper, we extend this study with sentiment analysis of video transcripts and by collecting and publishing a human-annotated ground truth dataset.

Beyond automatic emotion analysis, there naturally exists other types of research about the motivations and emotions of game streaming. For example, Sjöblom et al. [32] studied game spectating motivations using questionnaire measures, and Robinson et al. [28] designed technology for revealing the streamers' biometric data to the audience.

DATASET

This section describes the preparation of our dataset. We use the data for training and testing the automatic emotional event detection and classification system described in the next section. We also provide the dataset for other researchers as supplementary material.

Game and Video Selection

For our analysis we chose streams of the games *Unravel* [14] and its sequel, *Unravel Two* [15], developed by Swedish studio Coldwood Interactive and published by Electronic Arts for PC, PlayStation 4 and Xbox One. Both games are puzzle-platformers, where the player controls a humanoid figure made of yarn through different levels, overcoming obstacles and solving puzzles along the way. We chose these games for three reasons: First, they were recently released and readily featured on several Youtube channels. Second, the level design is largely linear, where all players experience game events in the same sequence. This ensured a certain consistency between streams and made for easier comparison. Third, and of particular interest for our analysis, both Unravel and Unravel Two were praised for being *emotionally* engaging [1, 2].

We browsed YouTube for streaming videos searching for "Unravel", "Unravel 2", and "Unravel Two". Videos were included in our analysis if they fulfilled the following criteria: (1) The streamer's face had to be visible throughout the video (e.g., not obstructed by a microphone). (2) The streamer provided commentary in English. (3) Only one person was playing and present during the stream. Hence, videos featuring local co-op (in Unravel Two) or where a person other than the streamer provided commentary were excluded. (4) Subtitle transcripts from automatic captioning had to be available for the video. Following this procedure, we selected 17 videos by 9 different streamers (2 women, 7 men), which encompass over 11 hours of video material. We only included videos that covered the first two levels of each game (i.e., "Thistle and Weeds" and "The Sea" in Unravel; "Foreign Shores" and "Hideaway" in Unravel Two). Refer to the supplementary material for the complete list of streams.

Emotion Expression Annotation

To establish ground truth, one of the authors first conducted an open coding of events depicting streamers' emotional expressions, where they reviewed all of the videos before manually developing a set of 18 initial research codes. Example event codes include "triumphant", "delighted when finding a collectable", "confused" and "insight/about to figure out a puzzle". We opted for a bottom-up open coding approach rather than applying a top-down coding framework, such as

Ekman's basic emotions [11, 12], because players' emotional expressions often go beyond basic emotions [18] and are not always clear-cut [29, 35]. In the Unravel videos, for instance, streamers often reacted emotionally to the "cuteness" of the main character(s), which has recently been argued to constitute its own distinct emotion [5]. Nevertheless, building upon previous work on emotional expression [11, 12, 18], the same author then manually collated the initial codes into a final set of 13 event codes, which were subsequently refined through discussion among the authors. Examples of the final event codes include streamers being "startled", "happy" or expressing "surprise", where each code was accompanied by a description of expressive features (e.g., streamers' eyes and mouth widening in surprise, as described by Ekman [12]). Full descriptions and examples of each event code are provided in the supplementary material. Note also that we included an event code "Not Applicable" (NA) for coding events that did not readily match any of the specified codes. This was done to take into account instances of emotional expression that were only rarely featured in the videos (e.g., the streamer wincing in pain as the main character crashes into a tree).

To ensure inter-rater reliability, four authors annotated the videos applying the final set of codes to identify emotionally salient events, whereby each video was coded by two authors. The reliability analysis is described in detail in the next section. Based on previous work on facial expression analysis in games [25, 29], we anticipated that accuracy of the automated expression annotation might be compromised for less expressive emotional responses. Hence, each annotator additionally flagged the Top 5 most intensely emotional events among all coded instances for each video, where we expected intense emotional events to also be accompanied by more pronounced emotional expression.

Data Windowing, Reliability, and Analysis Granularity

For formulating the automated emotion analysis as a standard classification problem, the data needs to be in the form of input feature vectors associated with output labels. To enable this, the data was divided into time windows with one second overlap, each window constituting a single data point. The input features for each window comprise various time series signals extracted from the video and audio. The output label of a window is either "no event" or the event category or categories found within the window.

In this paper, we perform all analyses with four different levels of granularity:

- 2 classes, either "no event" or "event". The latter class groups together all annotated events.
- 2 classes, only including the most intense 5 events each annotator identified for each video.
- 4 classes, including "no event", "pleasant event", "unpleasant event", and "neutral event". The latter three each correspond to a subset of the full 13 event codes.
- 14 classes, including "no event" and the full set of 13 codes.

As shown in Table 1, we tested different window lengths, which produce different inter-rater agreements, computed as the percentage of windows where both annotators either did not find any events or found an event and coded it to the same category. As short windows produce a large number of "no event" data points, we balanced the classes by weighting the data points to simulate equal data amounts for each class.

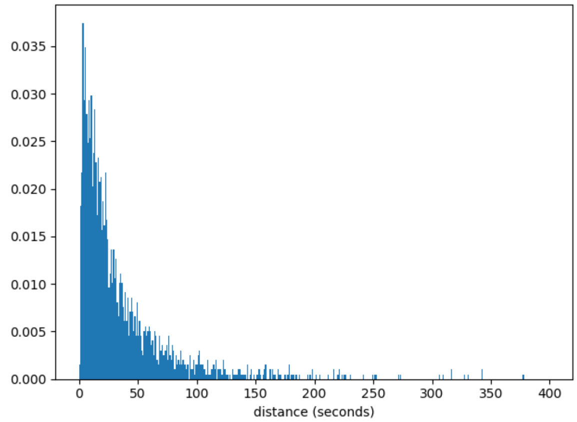

Figure 2. Histogram of distances between two consecutive events logged by the same annotator.

As one would expect, larger windows produce better inter-rater agreement, but also start to result in *congestion*, i.e., the same annotator logging multiple events per window. This signals insufficient temporal resolution. To minimize congestion to approximately 1% of windows, we limit our analysis to window lengths below 5 seconds. As shown in Figure 2, there are quite many events only a few seconds apart from each other, even though the average temporal distance between events is much higher.

Inter-rater agreement also decreases with finer granularity. Based on Table 1, it is clear that the 14-class data is not very reliable. On the other hand, inter-rater agreement is substantial in the 2-class case, ranging from 59.6% to 68.7% with windows from 1 to 5 seconds.

Finally, different annotators appear to have different sensitivity, some annotators logging more frequent events. In our data, the average distances between two consecutive events for different annotators range from 17 to 57 seconds. Although the selection of videos for each annotator also affects the annotation frequency, – where certain streamers were more expressive than others, – the variance is so high that individual differences in sensitivity probably play a role.

AUTOMATED MULTIMODAL EMOTION EXPRESSION ANNOTATION

This section describes our multimodal framework for automated emotion expression annotation. The system was implemented using the Keras deep learning framework [8]. As illustrated in Figure 1, our system comprises the following components:

- A facial expression analysis neural network that outputs probabilities of emotions such as "happy" or "surprised" based on each video frame.

Window length	Congestion	Inter-rater agreement			
		2-class	2-class-top events	4-class	14-class
1	0.0	59.6	53.4	34.9	18.8
2	0.1	64.3	56.8	39.3	24.0
3	0.3	67.0	59.1	41.8	27.3
4	0.5	68.3	61.0	43.1	29.4
5	1.1	68.7	60.3	43.7	30.8

Table 1. Inter-rater agreement and congestion with respect to different window lengths and levels of granularity

- An audio analysis module. We utilize a neural network that outputs similar emotional expression probabilities as the facial analysis, and we also extract additional low-level audio features such as loudness.
- A speech sentiment analysis neural network based on subtitle (speech transcription) data obtained from YouTube.
- A temporal convolutional neural network that receives windowed segments of the time series outputs of the modules above. This network is trained with our human-annotated dataset to output emotional event label probabilities. The training was done with four different levels of granularity, as explained in the previous section.

Obviously, as we have a fairly small dataset, we cannot train a single network end-to-end with raw video and audio as input and our annotations as training targets. Instead, we train the separate modules with existing datasets to extract relevant emotionally salient features. These features are combined by the final convolutional neural network trained with our own dataset.

In the subsections below, we give more detail of each module.

Facial Expression Analysis

We used the well-known VGG16 [31] convolutional neural network for predicting facial expression probabilities. The neural network is trained on Affectnet dataset [24]. This dataset contains around 500K manually annotated facial images labeled with emotions such as surprise, happy, and sad. Input images are resized to 48 × 48 pixels and normalized between zero and one. The last layer of the neural network uses the softmax activation function to map each image to emotion class. Data is downsampled to mitigate class imbalances; the downsampled training dataset contains 73806 images and the validation dataset has 18440 images. Validation accuracy was 60%. Figure 3 shows The confusion matrix.

Game streams usually include both gameplay video and facial video of the streamer. We used OpenCV [4] library for detecting faces in video frames. Detected faces are cropped, resized to 48 × 48 pixels, and pixel intensities are normalized to the range [0,1]. Sometimes, the face detection algorithm cannot find the face; the data in these frames is reconstructed using linear interpolation of the neighboring frames.

Audio Expression Analysis

For audio expression recognition, we again use a VGG16 image classification network, which is possible because we preprocess audio segments into 2D spectrogram images. We resample the audio signals at 16 KHz and compute power spectra using Short-time Fourier transform (STFT) with window size of 512 and stride 128. Power spectra are converted to decibel units, downsampled by a factor of 4, and normalized between zero and one. The Librosa library [21] was used for audio preprocessing. The power spectra of 3 seconds segments are concatenated to as a 64x94 pixel spectrogram image, which the VGG16 network maps to 7 classes of emotions. We used combination of different datasets ([6, 7, 10, 17, 19]) to train the network. In total, the number of training and validation data points is 9538 and 2384, respectively. The validation accuracy was 68%. Figure 4 shows the confusion matrix of the trained neural network.

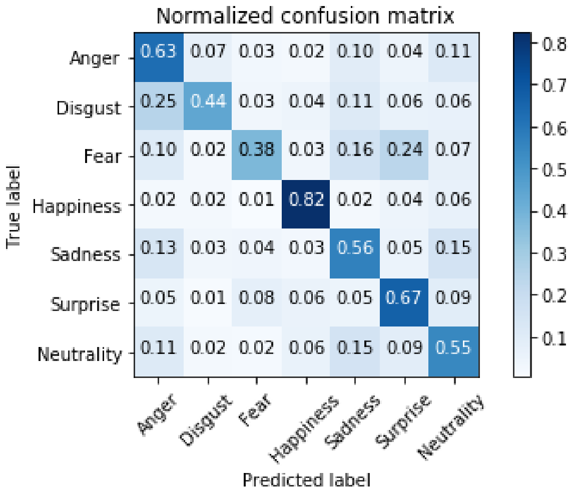

Figure 3. Confusion matrix of our VGG16 facial expression classification network.

A big problem with audio expression analysis is that game stream audio usually includes game music and sound effects in addition to the streamer's voice. This can cause errors especially when the player is not speaking. We mitigated this by detecting and discarding windows with no speech. This is done using the Librosa library [21] to remove the harmonic parts of the audio, which usually removes most of the music. We then discard the window if less than 30% of the remaining audio signal exceeds a magnitude threshold. The missing output values were reconstructed using linear interpolation. We also tried applying the audio emotion recognition to the signals with harmonic parts removed, but that yielded a lower classification accuracy.

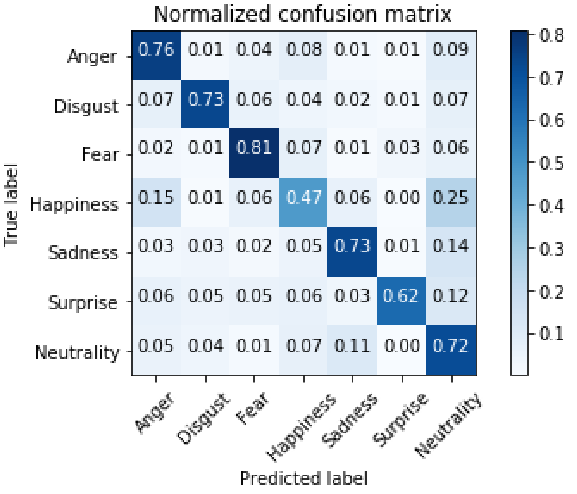

Figure 4. Confusion matrix of the audio expression analysis VGG16 network.

Audio features

The loudness and pitch of a streamers' voice often changes when they confront emotional events [27]. Therefore, we decided to also include pitch and loudness as additional features. The Librosa library was used to calculate root-mean-square audio signal power, pitch, and perceptually weighted loudness at each video frame.

Speech Transcript Sentiment Analysis

YouTube provides subtitles (transcripts based on speech recognition) which we use to analyze the streamers' speech. We use a convolutional neural network to process 3 second segments of the transcript data and predict the probability of positivity of what is said. We extract the segments around each video frame to output speech sentiment signals at the same frame rate as the facial expression probabilities. Naturally, a streamer does not speak all the time, in which case the sentiment output is linearly interpolated between nearest valid outputs.

The sentiment analysis neural network uses a combination of word embedding, temporal convolution, and a dense output layers (Table 2). A Kaggle dataset of Amazon reviews[1] is used for training. Sequences of 1000 words are used as training data. Longer sequences are truncated, and the ones with smaller length are padded with zero values. The neural network is trained on 3600k data and validated on 400k data. The validation accuracy is 94%.

Layer	Output shape
InputLayer	(batch size, 1000)
Embedding	(batch size, 1000, 100)
Conv1D	(batch size, 996, 128)
MaxPooling1D	(batch size, 199, 128)
Conv1D	(batch size, 195, 128)
MaxPooling1D	(batch size, 39, 128)
Conv1D	(batch size, 35, 128)
GlobalMaxPooling1D	(batch size, 128)
Dense	(batch size, 128)
Dense	(batch size, 2)

Table 2. The neural network architecture of speech sentiment analysis.

Layer	Output shape
InputLayer	(batch size, 300, 18)
Conv1D	(batch size, 298, 32)
MaxPooling1D	(batch size, 99, 32)
Conv1D	(batch size, 97, 32)
MaxPooling1D	(batch size, 32, 32)
Conv1D	(batch size, 30, 32)
MaxPooling1D	(batch size, 10, 32)
Conv1D	(batch size, 8, 32)
GlobalMaxPooling1D	(batch size, 32)
Dropout	(batch size, 32)
Dense	(batch size, n. of classes)

Table 3. The architecture of our final emotional event detection and classification network.

Mapping Features to Emotional Events

Finally, we employ a temporal 1D convolutional neural network that combines all the audio, video, and transcript features, processing them as multichannel time series data extracted from each analysis window. The network is trained to output the probabilities of our event codes. We use standard softmax cross-entropy classification loss function in the training. If an input data window contains multiple annotations, we add the window to the training data multiple times, once for each annotated label.

The network architecture is shown in Table 3. To achieve event detection invariant to temporal displacement of signals within the analysis windows, we employ a standard stack of 1D-convolution and max-pooling layers. As the receptive field of the neurons grows with the successive convolutions and max-poolings, the network is also better equipped to deal with possible delays between different input signals. The signals are normalized to have zero mean and unit standard deviation over each video. The normalization makes the network more invariant to the individual differences in emotional expressiveness of the streamers. Figure 5 shows an example of the input signals for the network within a 5 second window.

EVALUATION

As explained earlier, we carry out all analyses with four levels of granularity, ranging from 1 to 13 possible events in addition to "no event". All the granularities are handled similarly, except for the 5 most emotionally intense events recognition. In this case, there is very little training data, and we first trained with binary data (no event vs. event) and then froze all layers except the output layer, and finetuned the output layer with the top 5 event data. To investigate the effect of including multimodal input signals, we also tested the accuracy of our system both with all input types and with some input types disabled. Table 4 shows the validation accuracies and F_1-scores with different analysis window lengths and input signal

[1] https://www.kaggle.com/bittlingmayer/amazonreviews/discussion/3-3444

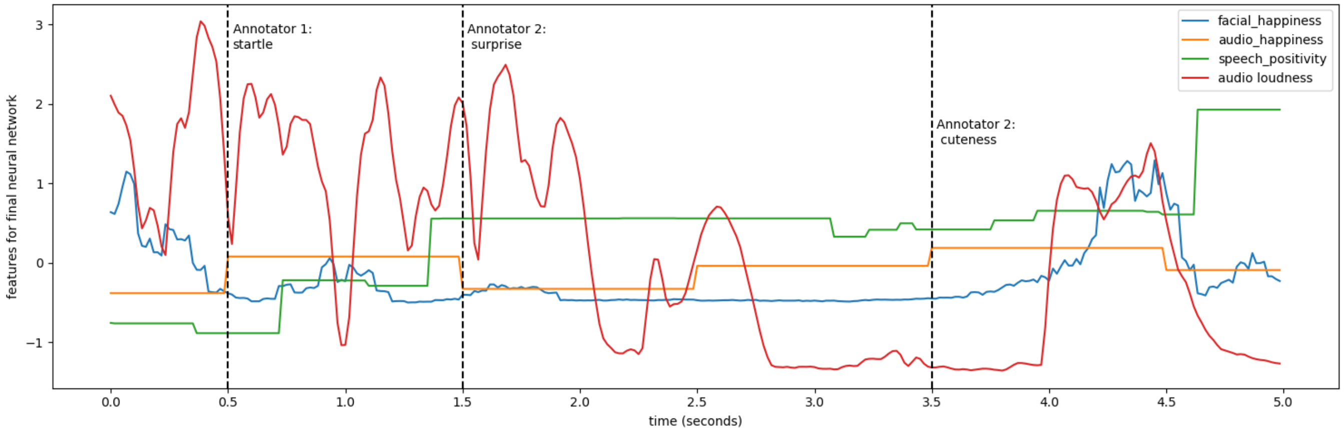

Figure 5. An example of the multimodal input signals of a 5 second window that the final network of our analysis pipeline maps to the annotated event probabilities, including the probability of no event. For clarity, only a subset of the signals are displayed. The figure also shows the annotations in our dataset. One sees how two annotators may disagree on the event codes such as "startle" and "surprise" and may log the event at slightly different times. Using large enough windows gets around the temporal inaccuracy, but increases the chance of congestion, i.e., multiple events logged in the same window by the same annotator.

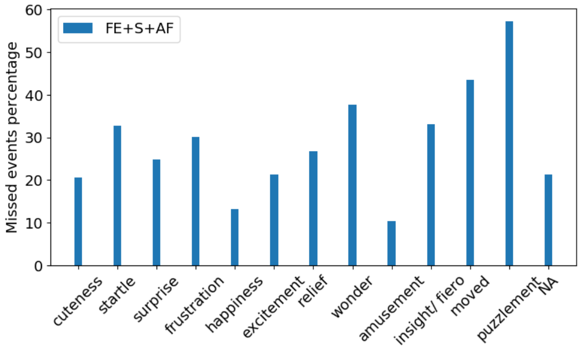

Figure 6. Percentage of missed events for each event category when using the best-performing 2-class (event vs. no event) classification based on facial expressions, speech sentiment and audio features.

combinations, with best scores for each window length shown in boldface. A randomly selected 20% subset of the data was used as the validation set. For increased reliability, the results were averaged over 5 independent training runs with different random seeds. Similar to the inter-rater agreements, the accuracies were computed with class imbalances corrected through training sample weighting. Due to the high number of data windows with no annotated events, the non-balance-corrected accuracies would be artificially high even if our system never detected any events. Likewise, F_1-scores were calculated using class weights to reduce the class imbalance bias.

The accuracy scores indicate that our system is the most useful for the binary classification case, i.e., detection of emotional events. Figure 6 shows the false negative rates for each of our 13 codes, i.e., how many events of each type one would miss in such a detection task, if using the best performing network that uses facial expressions, transcript sentiment analysis and audio features as the input. The figure indicates that the network is the most robust in detecting clearly positive emotions like amusement and happiness, but misses a large portion of puzzlement, moved, and wonder. The total numbers of validation set events used for creating Figure 6 are: cuteness: 37, startle: 84, surprise: 67, frustration: 78, happiness: 70, excitement: 59, relief: 14, wonder: 20, amusement: 70, insight/fiero: 132, moved: 9, puzzlement: 96, NA: 40.

DISCUSSION

The main insights that can be gained from investigating both our dataset and the evaluation results can be summarized as follows:

- Emotional event annotations can be temporally inaccurate and increasingly unreliable with more fine-grained classification. However, both our inter-rater reliability and automatic detection accuracy is reasonably good in basic detection of emotional events.
- With a limited number of classes, automated emotional event detection and classification is feasible and can produce human-level performance, i.e., accuracy similar to the inter-rater agreement. This is an encouraging result that should enable use cases like automated stream highlight recognition and pre-screening playtest videos for further analysis.
- At least in our case, facial expressions and audio features are the most informative signals, and audio expressions provide only limited improvement. However, this may be due to game music and sound effects corrupting the audio analysis. As a positive result that encourages future work, the confusion matrices of Figures 3 and 4 indicate that audio and facial expression analysis have complementary strengths.

Below, we further elaborate on these findings.

Granularity	Window length	Accuracy (%)					F_1-score (%)				
		FE	FE+S	FE+S+AE	FE+S+AF	Full	FE	FE+S	FE+S+AE	FE+S+AF	Full
2-class	1	63.0	63.5	63.6	**64.9**	**64.9**	60.0	60.3	61.4	**63.0**	62.3
	2	68.5	68.7	68.3	**70.7**	69.8	66.5	66.9	66.8	**69.9**	68.4
	3	67.9	67.5	67.0	**68.6**	67.5	65.8	65.2	64.3	**66.3**	64.3
	4	67.6	67.0	66.8	**68.7**	68.0	65.0	64.5	64.0	**66.4**	65.3
	5	68.1	67.6	67.1	**68.7**	68.0	65.1	65.0	64.0	**66.9**	65.1
2-class/ top events	1	70.2	68.7	67.3	**72.5**	71.3	69.1	66.8	64.6	**71.5**	69.3
	2	74.9	76.3	75.9	**80.4**	77.6	74.4	76.0	75.9	**80.7**	77.4
	3	74.3	73.4	73.4	75.6	**76.6**	73.2	71.7	71.6	74.1	**75.3**
	4	73.1	71.6	71.5	77.2	**78.0**	72.2	70.3	70.4	76.5	**78.0**
	5	71.2	69.0	71.3	74.2	**76.3**	69.0	65.8	69.1	72.6	**75.3**
4-class	1	41.9	**42.9**	42.2	40.3	39.5	51.8	**52.4**	51.9	49.5	49.1
	2	42.7	43.9	43.0	**44.5**	43.2	53.4	54.5	53.5	**55.4**	54.0
	3	**42.8**	42.3	40.3	42.5	41.6	**53.4**	53.1	50.6	53.3	52.1
	4	44.5	44.0	43.5	**45.4**	43.5	54.9	54.0	53.6	**55.5**	53.8
	5	41.0	41.7	41.9	**42.1**	41.7	51.1	51.8	**52.1**	52.0	51.8
14-class	1	19.8	**21.6**	21.0	19.4	20.7	29.5	**34.0**	33.1	29.2	32.0
	2	24.0	22.6	23.5	**26.4**	25.1	35.9	34.1	35.4	**39.4**	38.1
	3	18.3	**22.7**	21.3	21.7	21.3	28.1	**34.1**	32.2	32.9	31.6
	4	19.6	21.2	20.5	**23.8**	22.8	30.1	32.6	31.7	**34.9**	33.6
	5	19.7	19.1	19.7	**22.7**	20.8	29.5	29.3	30.7	**34.4**	32.2

Table 4. Accuracy and F_1-score of classification with different window lengths and levels of granularity. In each column, the final neural network has different inputs enabled. FE, S, AE, and AF denote facial expressions, speech (transcript) sentiment, audio expression analysis, and audio features, respectively. In the "full" column, all 4 types of inputs are used.

Human Annotation Accuracy

As shown in Figure 5, two annotators may log the same event with different emotion codes and timestamps. Naturally, there is more disagreement with more similar emotions such as being startled and surprised, in contrast to clearly distinct emotions like "angry" and "happy". Moreover, people tend to recognize intense emotional expressions more accurately than more subtle expressions [36]. Table 1 further reveals that the inter-rater agreement grows with window size, but at least in our data, windows of several seconds increase the chance of congestion. We have limited our analysis to congestion levels below 1%, which with our data translates to using windows of roughly 5 seconds or less. Please note also that if we had chiefly focused on human annotation (e.g., in the context of a qualitative study, where significant game events are defined and manually coded by the researchers [23, 33]), window size and congestion would not have been as relevant, and likely resulted in a higher inter-rater agreement.

Automated Analysis Accuracy

As shown in Figure 6, our system is most accurate at recognizing positive emotional expressions (cuteness, happiness, excitement, amusement, surprise). False negatives were more prevalent for subtle emotional expressions (e.g., puzzlement, moved, wonder), which may be due to lower intensity expressions being more difficult to recognize [36]. Moreover, expressions of frustration or being startled were frequently accompanied by immediate positive expressions (e.g., smiling), which may explain the relatively high amount of false negatives. This may be due to streamers expressing amusement after being startled, or as observed by previous work employing facial EMG [35] or automated facial expression analysis [29], players laughing off instances of failure (e.g., when missing a jump).

Table 4 shows that the accuracy of our system decreases with shorter time windows (better temporal resolution) and more fine-grained class detection. However, the detection and classification task is non-trivial for humans as well, and our accuracies are well in line with the inter-rater agreements. One notable difference is that with the automatic analysis accuracy, the relation to window length is not as pronounced. 1 second windows consistently produce worse results, but there appears to be no outstanding differences between the longer windows.

We find it interesting that our accuracy is particularly good in the top 5 most intense events detection. As such, even the 2-class/top events approach helps automate the detection of emotionally intense game events to some extent. For instance, our approach could be used to automatically summarize game stream highlights [27]. Alternatively, the system could be adapted for automatic identification of events in playtest videos, especially if the players are asked to think aloud while playing. These video highlights can then be shown to players to elicit additional feedback, as well as reduce unnecessary burden that might result from having players sit through an entire playtest recording. In contrast to previous work [23, 34, 35], however, our approach does not require specialized equipment or training in biometrics, which makes it more feasible for small and independent game developer studios [26].

Importance of Player Video vs. Audio

Based on Table 4, and in particular the best-performing 2-class cases, using multiple signals improves the results, compared to only analyzing facial expressions. However, the improvement is only a few percentages, and using all four input types only rarely leads to best results. Considering the fairly limited amount of training data, this may be due to the network overfitting to more high-dimensional data. To reduce overfitting, all added input variables should provide substantial additional information. Looking at Table 4, the low-level audio pitch and loudness features (AF) are such signals, as adding them often provides the best results.

From the point of view of analyzing the player's voice, the stream videos are corrupted by game music and sound effects. Table 4 suggests that in detecting the most emotional events, the audio emotion analysis (AE) is not very useful, whereas including the audio features (AF) does consistently improve the results.

On the other hand, we find it remarkable that with proper data, the same VGG16 neural network architecture provides almost similar accuracy in analyzing emotions from both audio spectrograms and images of a player's face. Comparing the confusion matrices in Figures 3 and 4, one sees that the networks have complementary strengths. Facial expression analysis is particularly good at recognizing happy expressions, whereas this is the primary weakness of the audio network. On the other hand, the audio network makes considerably fewer misclassifications of disgust, fear, and sadness. Thus, with non-corrupted data, audio and video could well complement each other and help disambiguate between emotion categories. Fortunately, recording only player speech without game audio can be easily done in non-streaming contexts by the player wearing headphones.

Limitations

In contrast to previous work analyzing players' emotional expressions (e.g., [23, 29, 34, 35]), we did not systematically take game events into consideration. First, this would have required a laborious and lengthy manual coding process, as the videos are not accompanied by logged and time-stamped game events (such as in [30]), or would require videos augmented with logged game event data (as in [16]). Second, we found that in our video sample, the same streamer's emotional reaction to the same event could vary substantially (e.g., due to being "distracted" by social interaction). On the other hand, not having to depend on knowledge on game events can also be considered a strength of our system. That said, adding automatic annotation of gameplay events from video (see [16]) as a complementary modality is a promising avenue for future work.

Another limitation is that we employed automatic captions for the text sentiment analysis. This was a necessary trade-off, as were interested in testing the utility of incorporating automatic modalities only, which allow for a relatively time-efficient and easy analysis. While text sentiment analysis would have likely been more accurate if we had manually transcribed streamers' utterances, it would have required a greater time investment.

Note also that we employed a fairly small dataset compared to previous work employing similar approaches (e.g., [29]). However, our results show that our approach reaches a satisfactory accuracy even with a limited dataset. This showcases its utility and generalizability for analyzing playtest video recordings and streams of a variety of games. As such, our approach is also suitable for small and independent game developer communities [26], who might otherwise not have access to games user research resources.

It should also be noted that our inter-rater agreement and validation accuracy scores are only approximately comparable. The inter-rater agreement would be more akin to the validation accuracy if only one annotator's data for each video was used for training and the other annotator's data was reserved for validation. However, this would mean that the exact same input features would be included in both the training and validation sets, making the validation accuracy not descriptive of overfitting. The approach would also discard a considerable amount of information about the variance of human annotator behavior. As we have relatively little data to start with, we instead used a random 80%-20% training and validation split.

Finally, our sample of streamers was rather homogeneous. All were Caucasian, based in Western Europe or North America, and likely aged in their twenties or early thirties. It remains to be seen whether our approach accounts for potential gender, age or cultural differences with regards to players' emotional expression. Given the limited number of streamers in our sample (2 women, 7 men), our analysis would not have yielded meaningful or reliable findings. However, seeing how previous work on automated emotion recognition has, for instance, observed culture-specific and cross-cultural nuances in facial emotion expressions [3], it is important for future work to account for the diversity of people who enjoy gaming and streaming.

CONCLUSION

We have presented a new dataset and automated detection and classification system for emotionally salient events in game stream videos. Our results indicate that identifying and classifying emotional events is a task that is hard for both humans and artificial neural networks. On the other hand, simplifying the task to only detecting the events yields a decent inter-rater agreement of 68.7%. Furthermore, our automated annotation system trained with the human annotations reaches a validation accuracy on par with the inter-rater agreement. Our system appears in particular usable for detecting the most intense emotional events, with accuracy of 80.4%, which suggests applications in automatic detection and summarization of video highlights or pre-selecting videos for further analysis after large-scale game testing.

The main technical novelty of our approach is that we utilize four different types of inputs and analyses: facial expression analysis, video transcript sentiment analysis, audio emotion analysis, and low-level audio feature extraction (pitch, loudness). Our evaluation indicates that it is indeed feasible to build such a multimodal neural network architecture with face, voice, and text analysis modules trained on existing large datasets, and finally combine the outputs of the modules using a network trained with a smaller custom dataset. Interestingly, our application of a VGG16 convolutional image classification network yields similar emotion recognition performance with both player facial images and voice spectrograms, with the face and voice classifications having complementary strengths. For example, fear is easily confused with other emotions based on the face, whereas it is the most robustly recognized emotion based on voice. On the other hand, this encouraging result based on the original voice emotion datasets does not necessarily generalize to game streams, where the audio is corrupted by game music and sound effects. This is probably one of the main reasons why we only see minor improvements when combining multiple input signals. In the future, we aim to test our approach with playtest videos recorded without game audio, with the players wearing headphones.

ACKNOWLEDGMENTS

We appreciate the valuable comments of anonymous reviewers. Hämäläinen has been supported by Academy of Finland grant 299358.

REFERENCES

[1] 2019a. *Unravel for PC Reviews - Metacritic*. (April 2019). https://www.metacritic.com/game/pc/unravel Retrieved April 5 2019.

[2] 2019b. *Unravel Two for PC Reviews - Metacritic*. (April 2019). https://www.metacritic.com/game/pc/unravel-two Retrieved April 5 2019.

[3] Gibran Benitez-Garcia, Tomoaki Nakamura, and Masahide Kaneko. 2018. Multicultural facial expression recognition based on differences of western-caucasian and east-asian facial expressions of emotions. *IEICE TRANSACTIONS on Information and Systems* 101, 5 (2018), 1317–1324. DOI: http://dx.doi.org/10.1073/pnas.1200155109

[4] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[5] Ralf C. Buckley. 2016. Aww: The Emotion of Perceiving Cuteness. *Frontiers in Psychology* 7 (2016), 1740. DOI:http://dx.doi.org/10.3389/fpsyg.2016.01740

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[7] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[8] François Chollet and others. 2015. Keras. https://keras.io. (2015).

[9] Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, and others. 2016. The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends® in Human–Computer Interaction* 9, 3-4 (2016), 151–308. DOI:http://dx.doi.org/10.1561/1100000065

[10] Kate Dupuis and M Kathleen Pichora-Fuller. 2010. *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department.

[11] Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion.. In *Nebraska symposium on motivation*. University of Nebraska Press.

[12] Paul Ekman. 2007. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan.

[13] Paul Ed Ekman and Erika L Rosenberg. 1997. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). (1997).

[14] Coldwood Interactive. 2016. *Unravel*. Game [PC, PlayStation 4, Xbox One]. (February 2016). Electronic Arts.

[15] Coldwood Interactive. 2018. *Unravel Two*. Game [PC, PlayStation 4, Xbox One]. (June 2018). Electronic Arts.

[16] Thanapong Intharah and Gabriel J Brostow. 2018. DeepLogger: Extracting User Input Logs From 2D Gameplay Videos. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM, 221–230.

[17] P Jackson and S Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).

[18] Nicole Lazzaro. 2009. Why we play: affect and the fun of games. *Human-computer interaction: Designing for diverse users and domains* 155 (2009), 679–700.

[19] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[20] Regan L Mandryk and Lennart E Nacke. 2016. Biometrics in Gaming and Entertainment Technologies. In *Biometrics in a Data Driven World*. Chapman and Hall/CRC, 215–248.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. 18–25.

[22] Elisa D. Mekler, Julia Ayumi Bopp, Alexandre N. Tuch, and Klaus Opwis. 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 927–936. DOI: http://dx.doi.org/10.1145/2556288.2557078

[23] Pejman Mirza-Babaei, Lennart E. Nacke, John Gregory, Nick Collins, and Geraldine Fitzpatrick. 2013. How Does It Play Better?: Exploring User Testing and Biometric Storyboards in Games User Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1499–1508. DOI: http://dx.doi.org/10.1145/2470654.2466200

[24] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* (2017).

[25] John T. Murray, Raquel Robinson, Michael Mateas, and Noah Wardrip-Fruin. 2018. Comparing Player Responses to Choice-Based Interactive Narratives Using Facial Expression Analysis. In *Interactive Storytelling*, Rebecca Rouse, Hartmut Koenitz, and Mads Haahr (Eds.). Springer International Publishing, Cham, 79–92.

[26] Lennart E. Nacke, Pejman Mirza-Babaei, Katta Spiel, Zachary O. Toups, and Katherine Isbister. 2018. Games and Play SIG: Engaging Small Developer Communities. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article SIG11, 4 pages. DOI:http://dx.doi.org/10.1145/3170427.3185360

[27] Charles Ringer and Mihalis A. Nicolaou. 2018. Deep Unsupervised Multi-view Detection of Video Game Stream Highlights. In *Proceedings of the 13th International Conference on the Foundations of Digital Games (FDG '18)*. ACM, New York, NY, USA, Article 15, 6 pages. DOI: http://dx.doi.org/10.1145/3235765.3235781

[28] Raquel Robinson, Zachary Rubin, Elena Márquez Segura, and Katherine Isbister. 2017. All the Feels: Designing a Tool That Reveals Streamers' Biometrics to Spectators. In *Proceedings of the 12th International Conference on the Foundations of Digital Games (FDG '17)*. ACM, New York, NY, USA, Article 36, 6 pages. DOI:http://dx.doi.org/10.1145/3102071.3102103

[29] Shaghayegh Roohi, Jari Takatalo, J. Matias Kivikangas, and Perttu Hämäläinen. 2018. Neural Network Based Facial Expression Analysis of GameEvents: A Cautionary Tale. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18)*. ACM, New York, NY, USA, 429–437. DOI:http://dx.doi.org/10.1145/3242671.3242701

[30] Noor Shaker and Mohammad Shaker. 2014. Towards Understanding the Nonverbal Signatures of Engagement in Super Mario Bros. In *User Modeling, Adaptation, and Personalization*, Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Springer International Publishing, Cham, 423–434.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[32] Max Sjöblom and Juho Hamari. 2017. Why do people watch others play video games? An empirical study on the motivations of Twitch users. *Computers in Human Behavior* 75 (2017), 985–996.

[33] Chek Tien Tan, Sander Bakkes, and Yusuf Pisan. 2014a. Correlation between facial expressions and the game experience questionnaire. In *Proceedings of the Entertainment Computing-ICEC 2014: 13th International Conference*, Vol. 8770. Springer, 229.

[34] Chek Tien Tan, Tuck Wah Leong, and Songjia Shen. 2014b. Combining Think-aloud and Physiological Data to Understand Video Game Experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 381–390. DOI: http://dx.doi.org/10.1145/2556288.2557326

[35] Wouter Van den Hoogen, Karolien Poels, Wijnand IJsselsteijn, and Yvonne de Kort. 2012. Between challenge and defeat: Repeated player-death and game enjoyment. *Media Psychology* 15, 4 (2012), 443–459. DOI:http://dx.doi.org/10.1080/15213269.2012.723117

[36] Tanja SH Wingenbach, Chris Ashwin, and Mark Brosnan. 2016. Validation of the Amsterdam Dynamic Facial Expression Set–Bath Intensity Variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions. *PloS one* 11, 1 (2016), e0147112.

[37] Georgios N Yannakakis and Ana Paiva. 2014. Emotion in games. *Handbook on affective computing* (2014), 459–471.