AI-Based Games User Research

Playability Evaluation, 26.10.2021

Christian Guckelsberger

(Slides adopted from Perttu Hämäläinen)

Self-Learning Agents Play Battlefield 1: https://www.youtube.com/watch?v=vaHnqXHQvF0 (SEED/Electronic Arts)

AI-Playtesting

- AI Playtesting: create an AI that can play a game.
- **Exercise**:
  What are use-cases?

What's the practical use of this technology right now?

Our short-term objective with this project has been to help the DICE team scale up its quality assurance and testing, which would help the studio to collect more crash reports and find more bugs.

In future titles, as deep learning technology matures, I expect self learning agents to be part of the games themselves, as truly intelligent NPCs that can master a range of tasks, and that adapt and evolve over time as they accumulate experience from engaging with human players.

https://www.ea.com/en-gb/news/teaching-ai-agents-battlefield-1

AI-Playtesting

- AI Playtesting: create an AI that can play a game.
- **Exercise**:
  What are use-cases?
  - Speed up playtesting
  - Explore levels more thoroughly
  - Find bugs / crash reports
  - Simulate different player types
  - Remove learning effects
  - …

https://medium.com/techking/human-like-playtesting-with-deep-learning-92adafffe921

Better Quality Content

A faster playtest allows for more iterations of the new levels. It means that level designers can refine more quickly. Because playtesting is not time-consuming anymore, it is possible to get feedback right before release to make sure that all tweaks work as intended. Finally, level designers can focus on the same content throughout the day, reducing the context switching mentioned above which impacts creativity.

More Thorough and Stable Playtests

One issue with human playtesters is that inherently, the more they play the better they get at the game. This introduces a bias into their feedback. Virtual players are version-controlled software, therefore avoiding such bias. On top of that, the measures are both more precise and diverse, since they communicate directly with the game engine.

A QA Byproduct

By building an automated playtesting platform for content balancing purposes, we actually created a QA byproduct for developers. They can use the platform to explore levels and find bugs. They can also check that new features don't break the rest of the game. It is a powerful tool to increase the game's quality as a whole.

~~AI-Playtesting~~
AI-Based Games User Research

- But many more uses of AI in Games-User Research!
  - E.g. unsupervised learning (k-means clustering) to identify player types from gameplay features.
- And often a combination of techniques:
  - E.g. reinforcement learning for AI-playtesting + linear regression to predict experience from gameplay features.
- Many different forms, not standardised

Predicting Game Difficulty and Engagement Using AI Players

SHAGHAYEGH ROOHI, Aalto University, Finland
CHRISTIAN GUCKELSBERGER, Aalto University, Finland
ASKO RELAS, Rovio Entertainment, Finland
HENRI HEISKANEN, Rovio Entertainment, Finland
JARI TAKATALO, Rovio Entertainment, Finland
PERTTU HÄMÄLÄINEN, Aalto University, Finland

This paper presents a novel approach to automated playtesting for the prediction of human player behavior and experience. It has previously been demonstrated that Deep Reinforcement Learning (DRL) game-playing agents can predict both game difficulty and player engagement, operationalized as average pass and churn rates. We improve this approach by enhancing DRL with Monte Carlo Tree Search (MCTS). We also motivate an enhanced selection strategy for predictor features, based on the observation that an AI agent's best-case performance can yield stronger correlations with human data than the agent's average performance. Both additions consistently improve the prediction accuracy, and the DRL-enhanced MCTS outperforms both DRL and vanilla MCTS in the hardest levels. We conclude that player modelling via automated playtesting can benefit from combining DRL and MCTS. Moreover, it can be worthwhile to investigate a subset of repeated best AI agent runs, if AI gameplay does not yield good predictions on average.

CCS Concepts: • **Human-centered computing** → **User models**; • **Computing methodologies** → *Modeling and simulation.*

Additional Key Words and Phrases: Player Modelling, AI Playtesting, Game AI, Difficulty, Player Engagement, Pass Rate Prediction, Churn Prediction, Feature Selection

ACM Reference Format:
Shaghayegh Roohi, Christian Guckelsberger, Asko Relas, Henri Heiskanen, Jari Takatalo, and Perttu Hämäläinen. 2021. Predicting Game Difficulty and Engagement Using AI Players. *Proc. ACM Hum.-Comput. Interact.* 5, CHIPLAY, Article 231 (September 2021), 18 pages. https://doi.org/10.1145/3474658

1 INTRODUCTION

The development of a game typically involves many rounds of playtesting to analyze players' behaviors and experiences, allowing to shape the final product so that it conveys the design intentions and appeals to the target audience. The tasks involved in human playtesting are repetitive and tedious, come with high costs, and can slow down the design and development process substantially. Automated playtesting aims to alleviate these drawbacks by reducing the need for human participants [10, 22, 56]. An active area of research, it combines human computer interaction

AI-Based Games User Research

- Complement + Augment GUR
  - E.g. to save time/cost in manual playtesting (augment)
  - Or to enable player experience / behaviour prediction on new levels (complement)
- **Exercise**: What are the main risks?
- Use traditional GUR techniques as:
  - data source
  - to validate whether AI method works
  - …

Predicting Player Experience without the Player An Exploratory Study

Christian Guckelsberger[1,*], Christoph Salge[2,3], Jeremy Gow[1], and Paul Cairns[4]

[1]Computational Creativity Group, Goldsmiths, University of London, London, UK
[2]Adaptive Systems Research Group, University of Hertfordshire, Hatfield, UK
[3]Game Innovation Lab, New York University, New York, US
[4]Department of Computer Science, University of York, York, UK
[*]Corresponding author. Email: c.guckelsberger@gold.ac.uk

ABSTRACT

A key challenge of procedural content generation (PCG) is to evoke a certain player experience (PX), when we have no direct control over the content which gives rise to that experience. We argue that neither the rigorous methods to assess PX in HCI, nor specialised methods in PCG are sufficient, because they rely on a human in the loop. We propose to address this shortcoming by means of computational models of intrinsic motivation and AI game-playing agents. We hypothesise that our approach could be used to automatically predict PX across

PCG algorithms require formal guidelines about the desire content characteristics. A procedurally generated level shoul without doubt be *playable*, i.e. there must be a way for th player to succeed or fail, or to experience the whole conter instance and not just a small part of it. Content should als be *novel* and *typical* (cf. [38]): a generated quest for instanc should be different from existing quests, but still fit the gam under consideration. However, nobody would care about level, character or as a consequence even the overall game, the content in question did not lead to a desired experienc

Session 5: Tools to Analyse Games CHI PLAY 2017, October 15–18, 2017, Amsterdam, NL

Figure 3: The *RoboRunner* testbed: a deterministic, one-button (in)finite runner game.

$$E[\mathfrak{E}^n]_{s_t} = \sum_{a_t} p(a_t|s_t) \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t)\, \mathfrak{E}^n_{s_{t+1}} \qquad (2)$$

This definition accounts for the possibility of noise in the agent's local dynamics, i.e. the player's forward model: a player might not be sure about the consequences of their actions, or the action outcomes might objectively be uncertain. Figure 1 shows the two stages in the calculation of 3-step state-expected empowerment at time t: the agent first anticipates which states its actions might yield at $t+1$ (- -), and

sponse of players by providing them different games to play, but the data is analysed qualitatively in order to explore the concepts in play around the manipulation. For this reason, only modest numbers of participants are required. Different conditions here are given by different level instances of an (in)finite runner game. Our hypothesis is that levels with low mean state-expected empowerment evoke qualitatively different experiences than levels with high values. We conduct a *thematic analysis* [6] on player think-alouds to find out which experiences the different conditions give rise to. We decided against a more quantifiable approach such as content analysis,

AI-Playtesting

”Attempting to maximize coverage of a game viahuman gameplay is laborious and repetitive, introducing delays in the development process. Despite the importance of quality assurance (QA) testing, QA remains an underinvested area in the technical games research community. In this paper, we show that relatively simple automatic exploration techniques can be used to multiplicatively amplify coverage of a game starting from human tester data”.

Reveal-More: Amplifying Human Effort in Quality Assurance Testing Using Automated Exploration

Kenneth Chang
University of California, Santa Cruz
Santa Cruz, CA, USA
kchang44@ucsc.edu

Batu Aytemiz
University of California, Santa Cruz
Santa Cruz, CA, USA
baytemiz@ucsc.edu

Adam M. Smith
University of California, Santa Cruz
Santa Cruz, CA, USA
amsmith@ucsc.edu

***Abstract*—Attempting to maximize coverage of a game via human gameplay is laborious and repetitive, introducing delays in the development process. Despite the importance of quality assurance (QA) testing, QA remains an underinvested area in the technical games research community. In this paper, we show that relatively simple automatic exploration techniques can be used to multiplicatively amplify coverage of a game starting from human tester data. Instead of attempting to displace human QA efforts, we seek to grow the impact that a human tester can make. Experiments with two games for the Super Nintendo Entertainment System highlight the qualitative and quantitative differences between isolated human and machine play compared to our hybrid approach called Reveal-More. We contribute a QA testing workflow that scales with the amount of human and machine time allocated to the effort.**

I. Introduction

In quality assurance (QA) testing for videogames, conventional wisdom holds that automated approaches answer *software* questions (e.g. does processing this sequence of inputs yield the expected output?) and manual testing answers *gameplay* questions (e.g. will the game crash if I collect this item?). Nascent research efforts in automatic testing have tried to apply artificial intelligence (AI) methods to the problem of demonstrating interesting possibilities in play that developers might interpret to answer design and implementation questions that impact gameplay. So far, separated human and machine testing processes have shown complementary strengths [1], as expected [2]. In this paper, we are interested in directly amplifying human tester effort to answer gameplay questions by using recordings of their play as the seeds for automated exploration.

Without automation, identifying inputs that lead to gameplay issues is a massive exploratory search problem that requires significant resource expenditure. Even in the simplest of videogames, there may be an astronomical number of distinct gameplay paths, only a few of which trigger a bug. In an ideal world, QA testers would indicate which span of a game is most relevant to them, and a system would quickly show them what was possible (or impossible) in that part of the game. Testers would save their efforts for directing, rather than enacting, repetitive gameplay experiments. Towards this goal, we formulate our problem as maximizing game state coverage in the service of encountering game design problems.

978-1-7281-1884-0/19/$31.00 ©2019 IEEE

While there has been high profile successes in automatic gameplaying research [3], only recently has exploration specifically drawn attention [4]. Score optimization techniques such as Reinforcement Learning (RL) [5] and Monte-Carlo Tree Search (MCTS) [6] are setup to solve a different problem from the one faced in exploration. Techniques like MCTS may systematically avoid exploring certain play styles of interest simply because they earn lower scores. Additionally, the timescale on which automated gameplay techniques achieve useful results (i.e. minutes versus years of simulated gameplay) has only recently drawn attention [4]. For exploration to be useful in the QA process, useful reports need to be generated on timescales comparable to the pace of game design cycles (such as being able to provide feedback on weekly or daily game builds).

In this paper, we demonstrate a new technique, Reveal-More, that combines automatic exploration with just minutes of human gameplay, resulting in game state coverage that is superior to using each individual method alone. In such a manner, an automated method of exploration is used to amplify what a person can contribute to testing, thus lowering the strain placed upon testers to find all the paths in a game. To anchor our work in game development practice, we carry out experiments in the commercial implementation of two culturally significant games. In several experiments with *Super Mario World* and *The Legend of Zelda*, we demonstrate up to a 5X increase in our quantitative exploration metric, and qualitatively illustrate the significance of increased coverage. Furthermore, we show that this amplified coverage can be helpful in visualizing design changes and, in turn, help characterize the impact of design changes.

II. Related Work

Common practice in game QA testing involves having many people play the game with the goal of covering the most ground in it. There exists some automation towards this goal [7], however the majority of the technical games research community has focused on creating algorithms that aim to maximize in-game score. In the search for the best QA practices, whether through automation or manual testing, many have agreed that maximizing some sense of *coverage* is a central concern [8]–[10].

AI-Playtesting

Reveal-More: Amplifying Human Effort in Quality Assurance Testing Using Automated Exploration

Kenneth Chang
University of California, Santa Cruz
Santa Cruz, CA, USA
kchang44@ucsc.edu

Batu Aytemiz
University of California, Santa Cruz
Santa Cruz, CA, USA
baytemiz@ucsc.edu

Adam M. Smith
University of California, Santa Cruz
Santa Cruz, CA, USA
amsmith@ucsc.edu

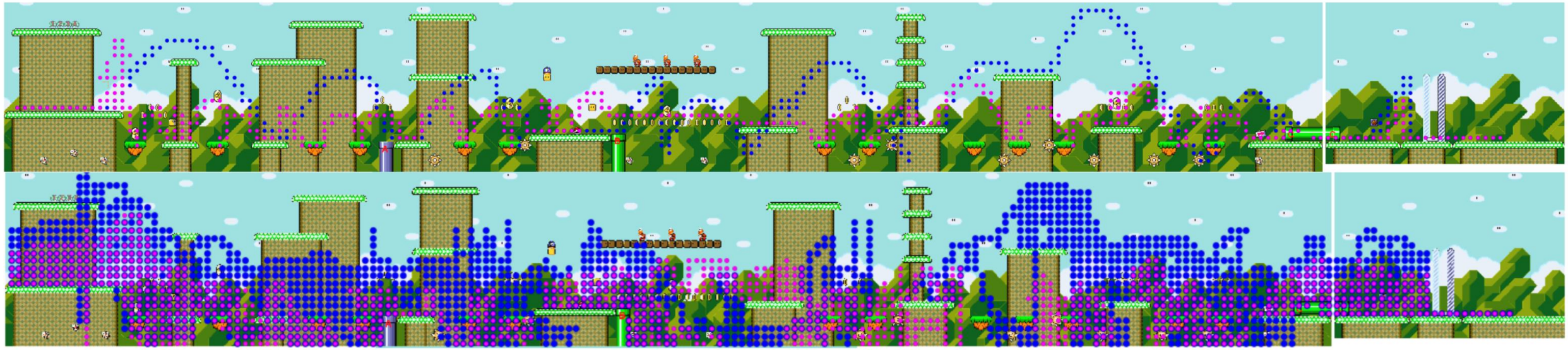


Fig. 8: A comparison of two different versions of SMW. The top image shows the human gameplay trace in the original (magenta) and modified (blue) designs while the bottom image shows the amplified coverage discovered with Reveal-More.

techniques can be used to multiplicatively amplify coverage of a game starting from human tester data".

Without automation, identifying inputs that lead to gameplay issues is a massive exploratory search problem that requires significant resource expenditure. Even in the simplest of videogames, there may be an astronomical number of distinct gameplay paths, only a few of which trigger a bug. In an ideal world, QA testers would indicate which span of a game is most relevant to them, and a system would quickly show them what was possible (or impossible) in that part of the game. Testers would save their efforts for directing, rather than enacting, repetitive gameplay experiments. Towards this goal, we formulate our problem as maximizing game state coverage in the service of encountering game design problems.

978-1-7281-1884-0/19/$31.00 ©2019 IEEE

helpful in visualizing design changes and, in turn, help characterize the impact of design changes.

II. Related Work

Common practice in game QA testing involves having many people play the game with the goal of covering the most ground in it. There exists some automation towards this goal [7], however the majority of the technical games research community has focused on creating algorithms that aim to maximize in-game score. In the search for the best QA practices, whether through automation or manual testing, many have agreed that maximizing some sense of *coverage* is a central concern [8]–[10].

Emotion Analysis

- Here focus on one aspect: emotion analysis of players
- What can we do with it?
- For (i) wide range of applications and (ii) straight-forward applicability in your GUR and ARTS projects!

CHALLENGE TYPES

Cognitive Challenge: Challenge that addresses the player's cognitive and problem-solving capacities. The player has to invest cognitive effort to predict the consequences of actions or comprehend ambiguous elements of the narrative or the storyline.

Physical Challenge: Challenge that addresses the player's physical limitations to interact with the game, i.e. the speed and accuracy with which actions can be performed.

Emotional Challenge: Challenge which confronts the player with emotionally salient material or the use of strong characters, and a captivating story. A player cannot overcome emotional challenge with skill or dexterity, but by resolving tension in the narrative, by identifying with characters, and by resolving ambiguities.

ELSEVIER

International Journal of Human-Computer Studies

Volume 137, May 2020, 102383

Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS)

Alena Denisova [a] ✉, Paul Cairns [b], Christian Guckelsberger [c], David Zendle [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.ijhcs.2019.102383

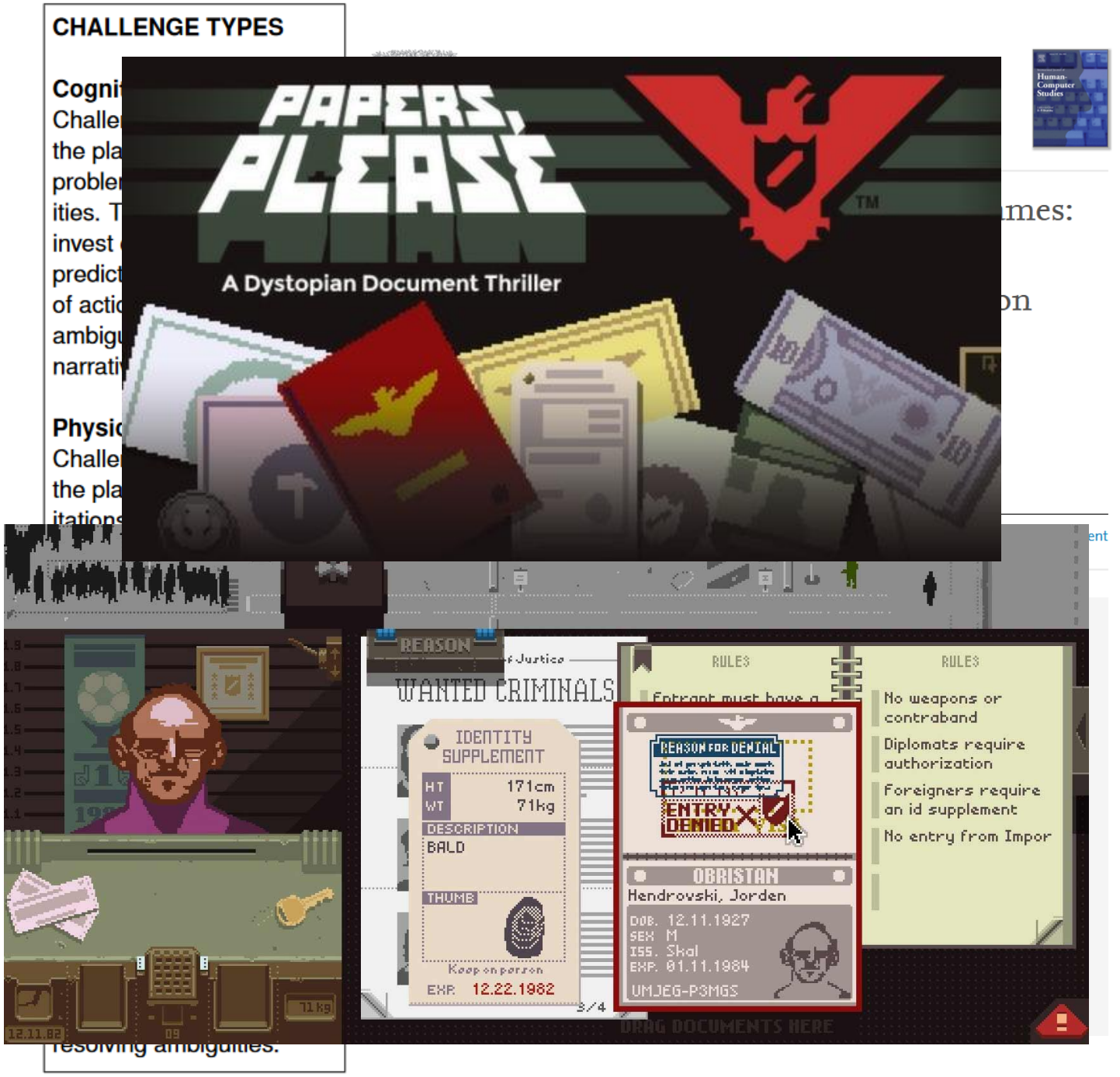
Get rights and content

Highlights

- Scale measuring perceived challenge in video games.
- Four sub-scales (30 items) measuring four types of perceived challenge in video games: cognitive, performative, emotional, and decision-making challenge.
- Development and validation are carried out over three studies including 1390 players with diverse backgrounds playing video games from a range of genres.
- The questionnaire is a systematic, extensive, reliable, and valid tool to measure perceived challenge in video games.

Emotion Analysis

- Here focus on one aspect: emotion analysis of players
- What can we do with it?
- For (i) wide range of applications and (ii) straight-forward applicability in your GUR and ARTS projects!

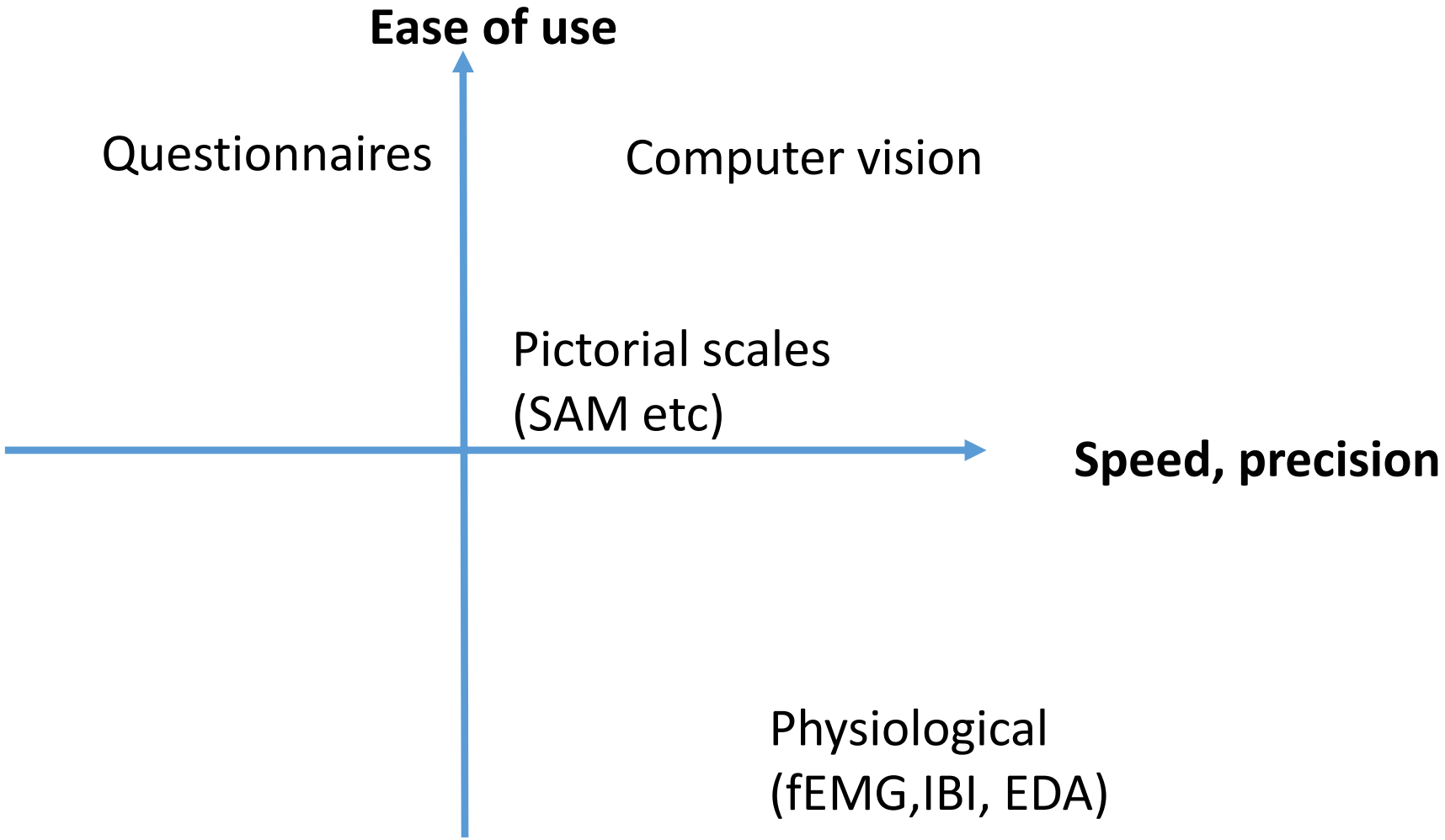


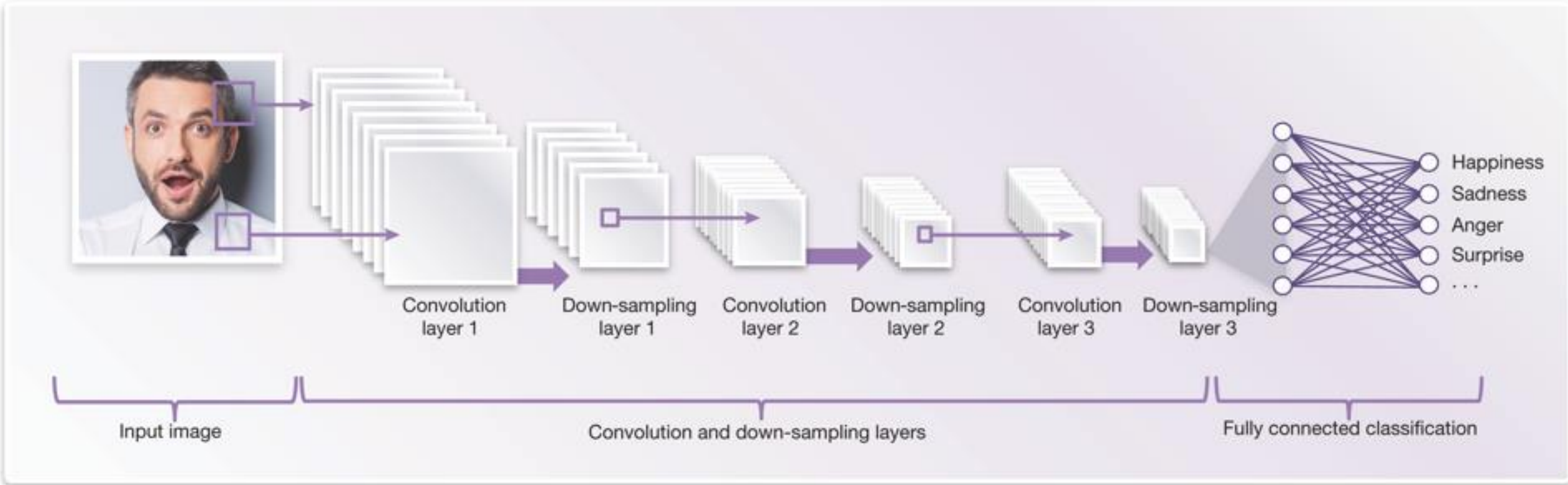
Use-Case:
Scalable emotion analysis from video, audio & text

Prof. Perttu Hämäläinen

With additions by Dr Christian Guckelsberger

Measuring emotion



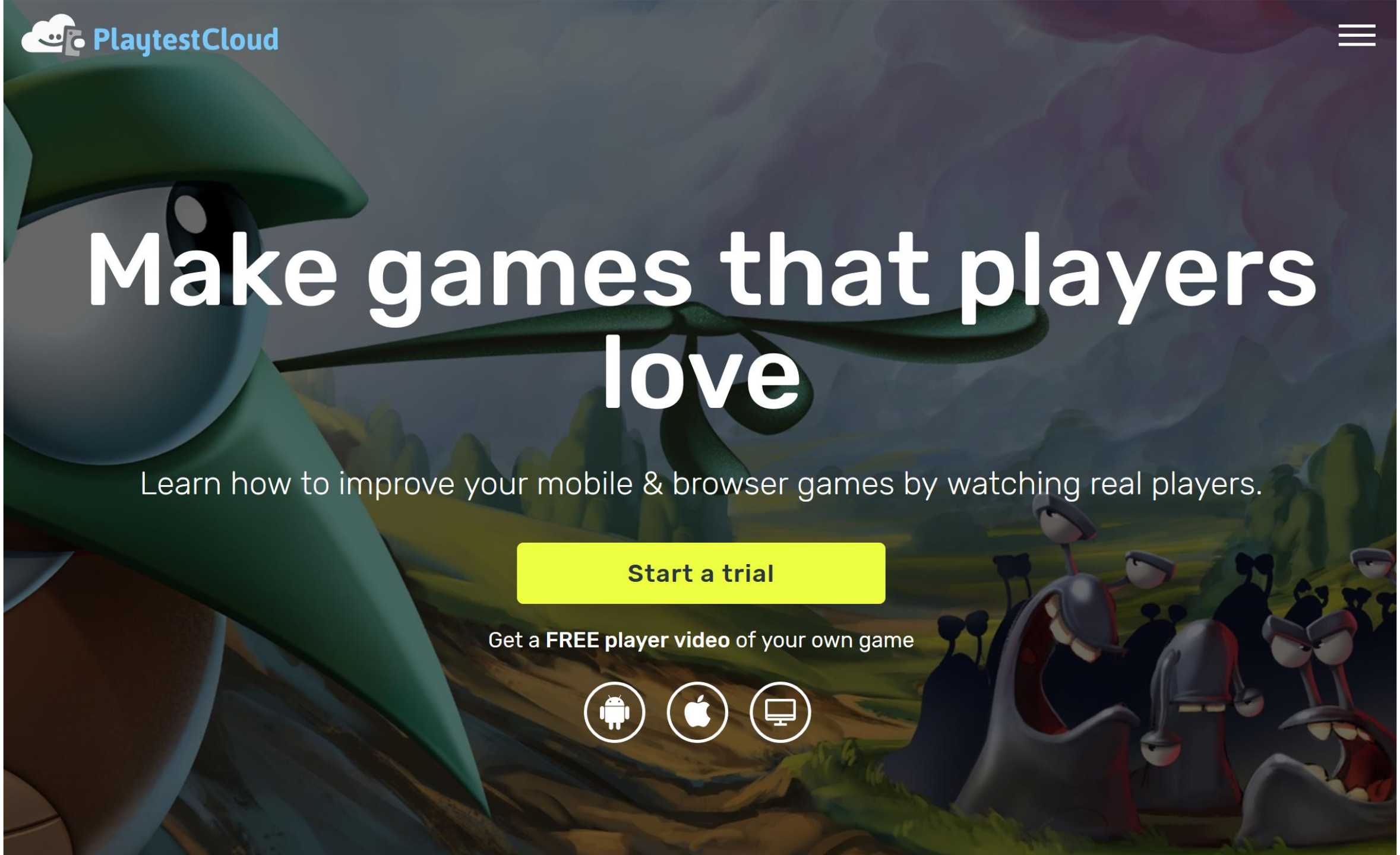


http://www.techdesignforums.com/practice/technique/facial-recognition-embedded-vision/


ABIERTO
“joyLikelihood”: “VERY_LIKELY”
,”description”: ”ABIERTO\n”,
“local”: “es”

facial electromyography (fEMG)

Image Source: http://www.emg-eeg.com/tag/beyin-damar-ultrasonu


PlaytestCloud
Make games that players love
Learn how to improve your mobile & browser games by watching real players.
Start a trial
Get a FREE player video of your own game

PlaytestCloud

Products

Surveys

starting at **$9** /response

Send a survey to mobile gamers
from any target audience.

Playtesting

starting at **$49** /player

First-time player experience testing.
Players play 5 minutes or longer.

Multi-Session

starting at **$900** /playtest

Optimize the player experience
in the first 2-5 gameplay sessions.

Longitudinal

starting at **$1,020** /study

Test the first 3-10 days of gameplay.
Players play one or more sessions per day.

Neural Network Based Facial Expression Analysis of Game Events: A Cautionary Tale

Shaghayegh Roohi, Jari Takatalo, J. Matias Kivikangas, Perttu Hämäläinen

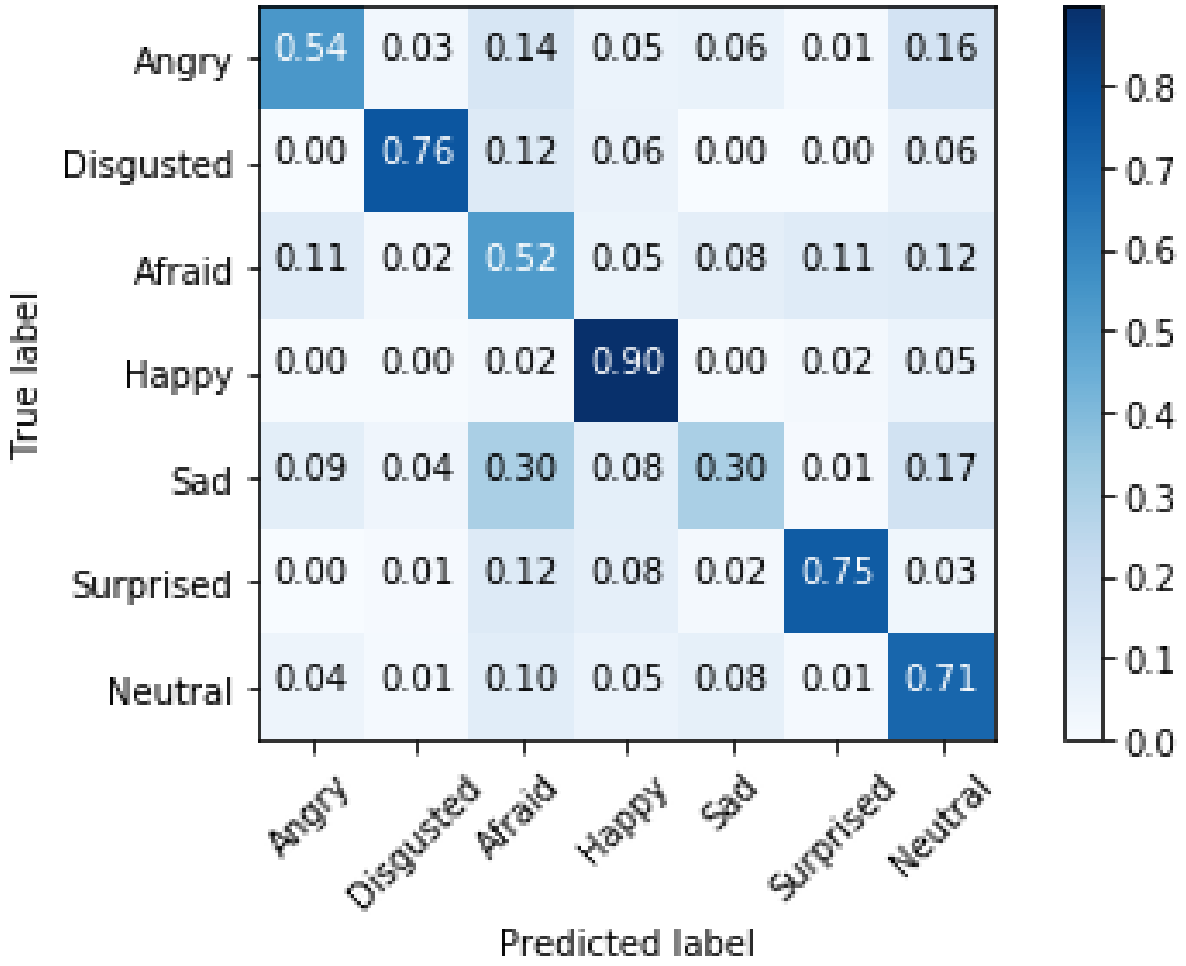
Decision in ”Life Is Strange”. https://rememberme.fandom.com/wiki/Choice_and_Consequence

Why focus on events?

- Every player may take a different path through a game
- Hence: Time-series signals of emotions cannot be directly compared or aggregated
- However, it’s easy to log out key events with time stamps: player decision, player getting a reward etc.
- Here: measuring the **affect gradient** of events: Average change in emotional facial expression around specific events.

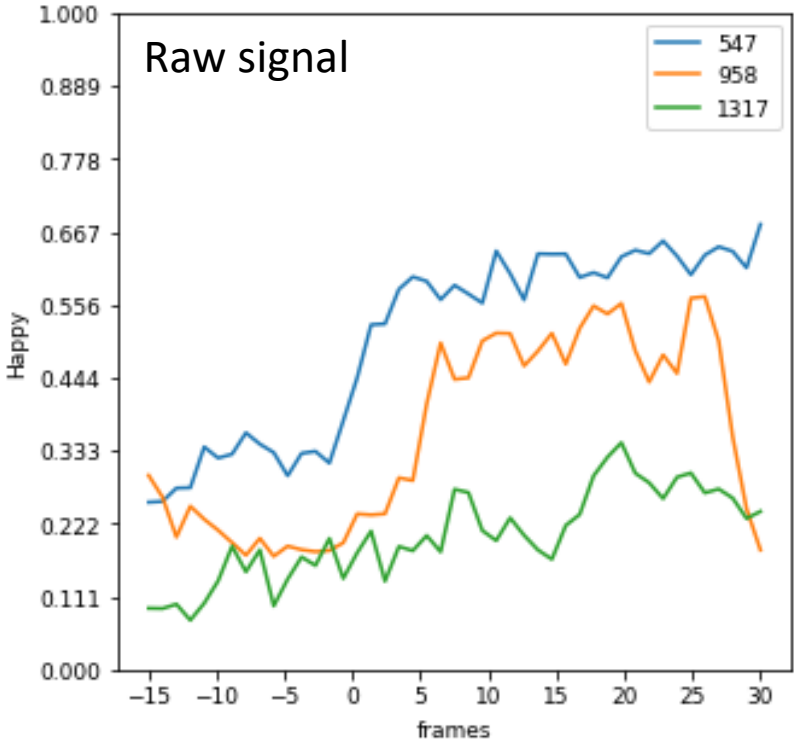
Approach

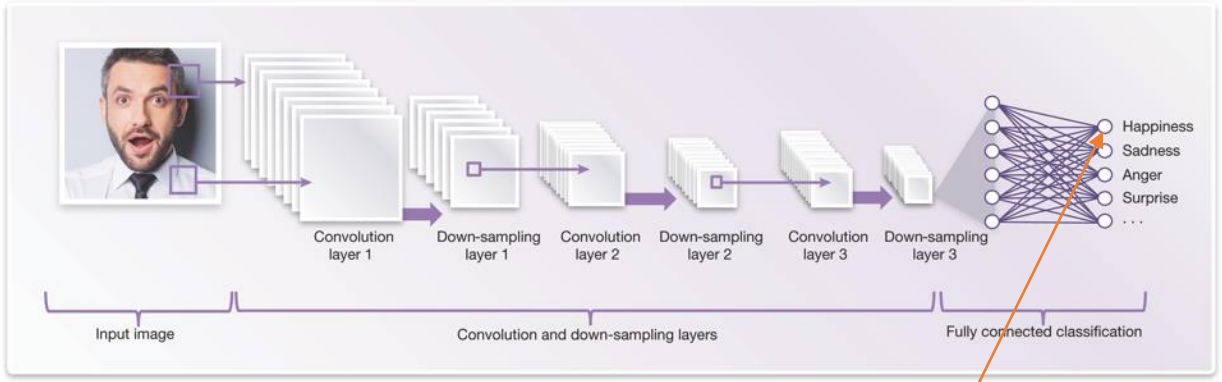
- Computer vision using facial video of the player
- Convolutional neural networks trained with free 7-emotion Kaggle dataset
- Focus on ”Happy”, because it’s most robust
- Analyze response to game events like dying or killing enemy



Affect gradient

- X: frames before/after event
- Y: class probability

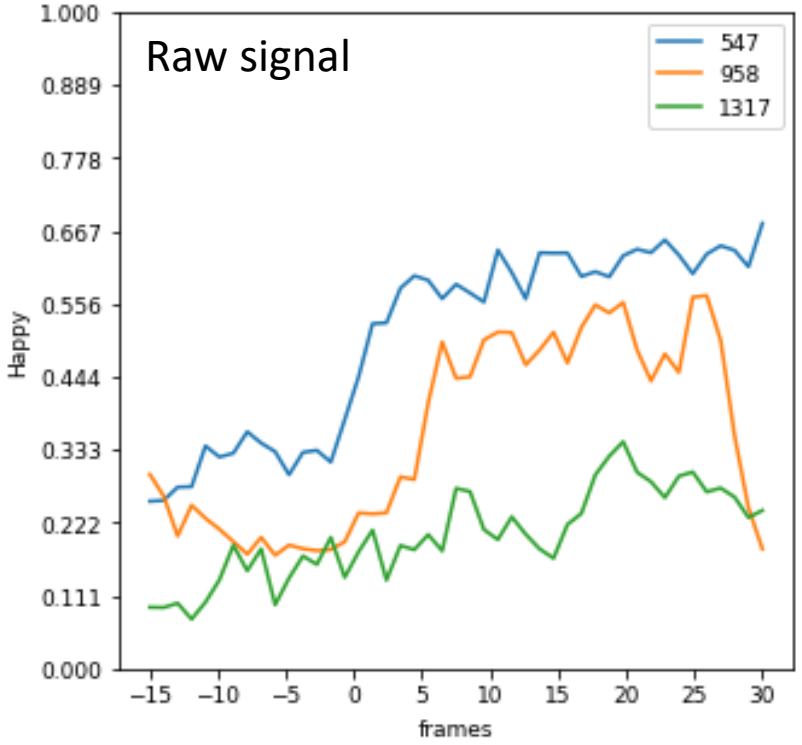


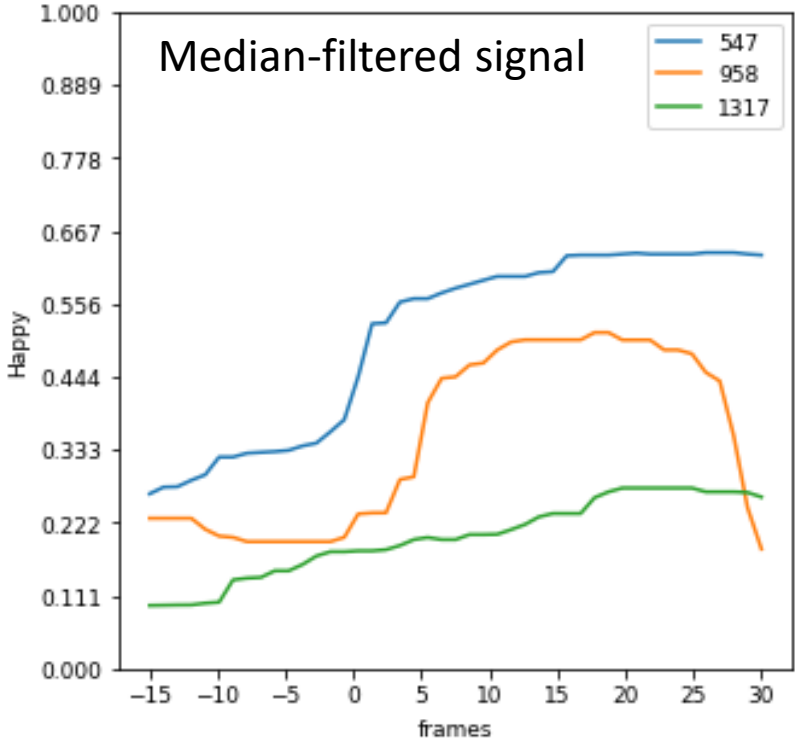


p(Happyness) in [0,1]

Affect gradient

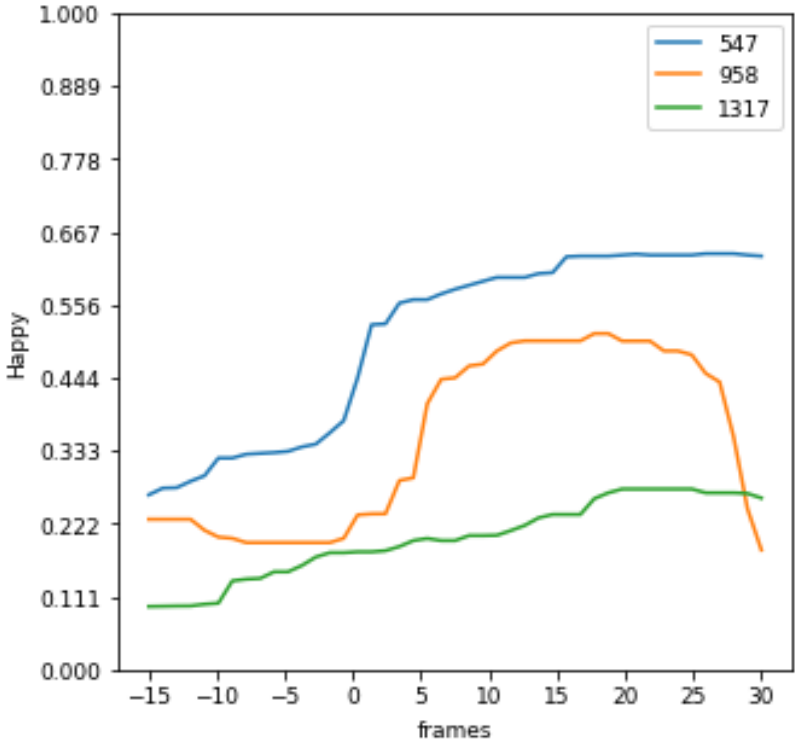
- Preprocess the data with median filtering
- Extract segments of fixed length around each logged game event



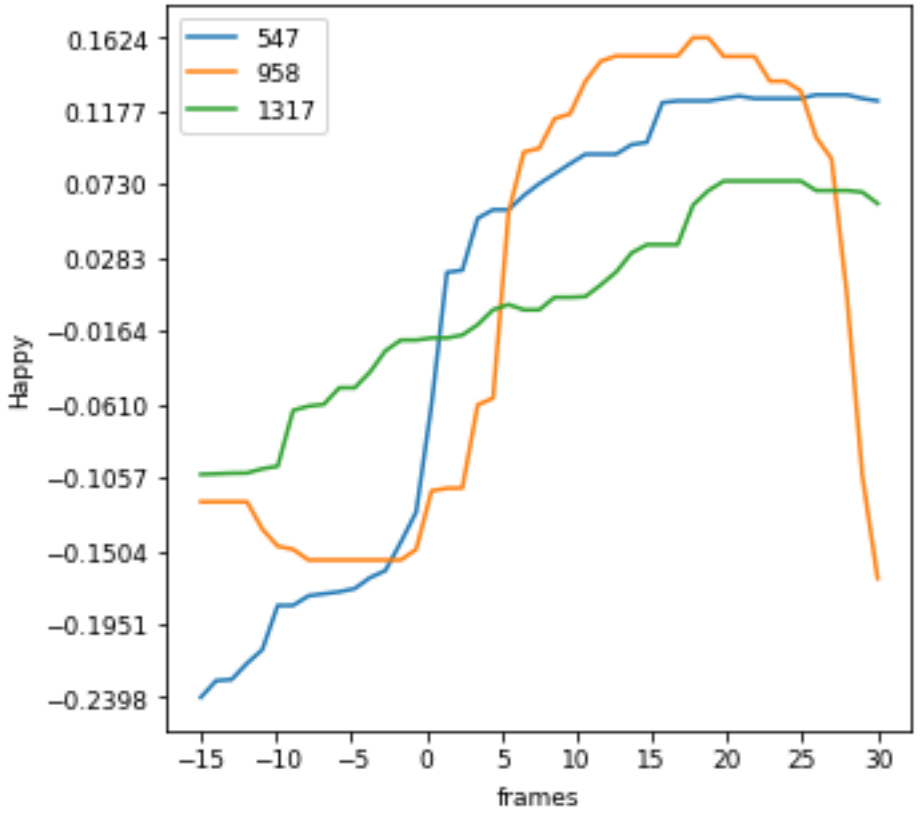


Affect gradient

- Normalize the segments to have zero mean



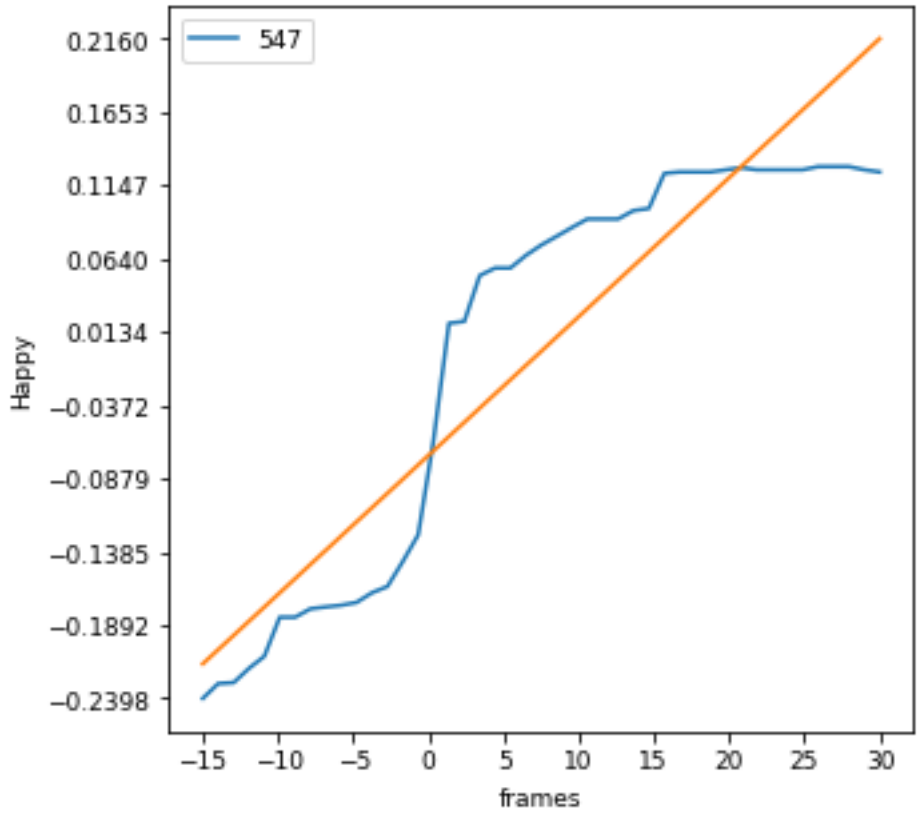
Median-filtered signal



Normalized signal

Affect gradient

- Fit lines to the extracted segments
  - The slopes give the affect gradients of the events

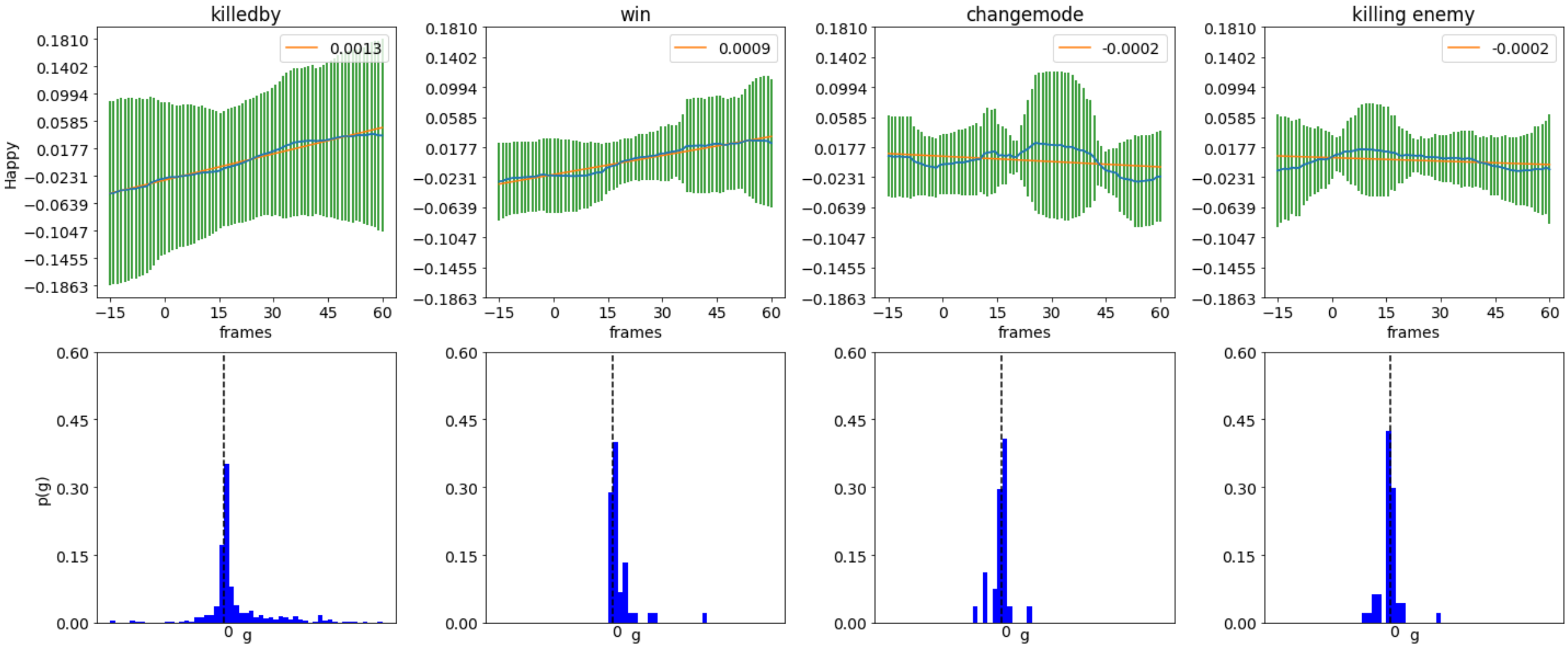


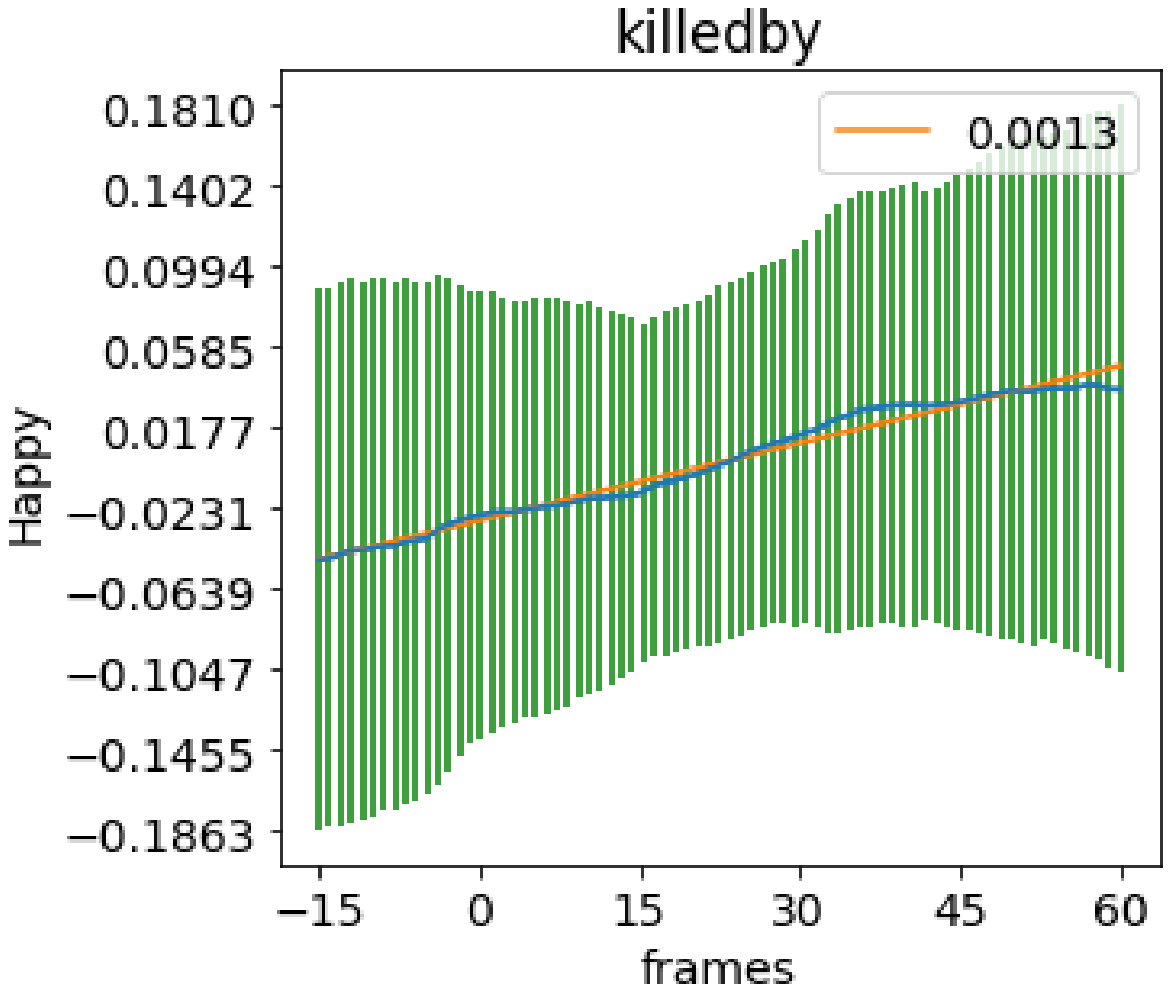
Affect gradient

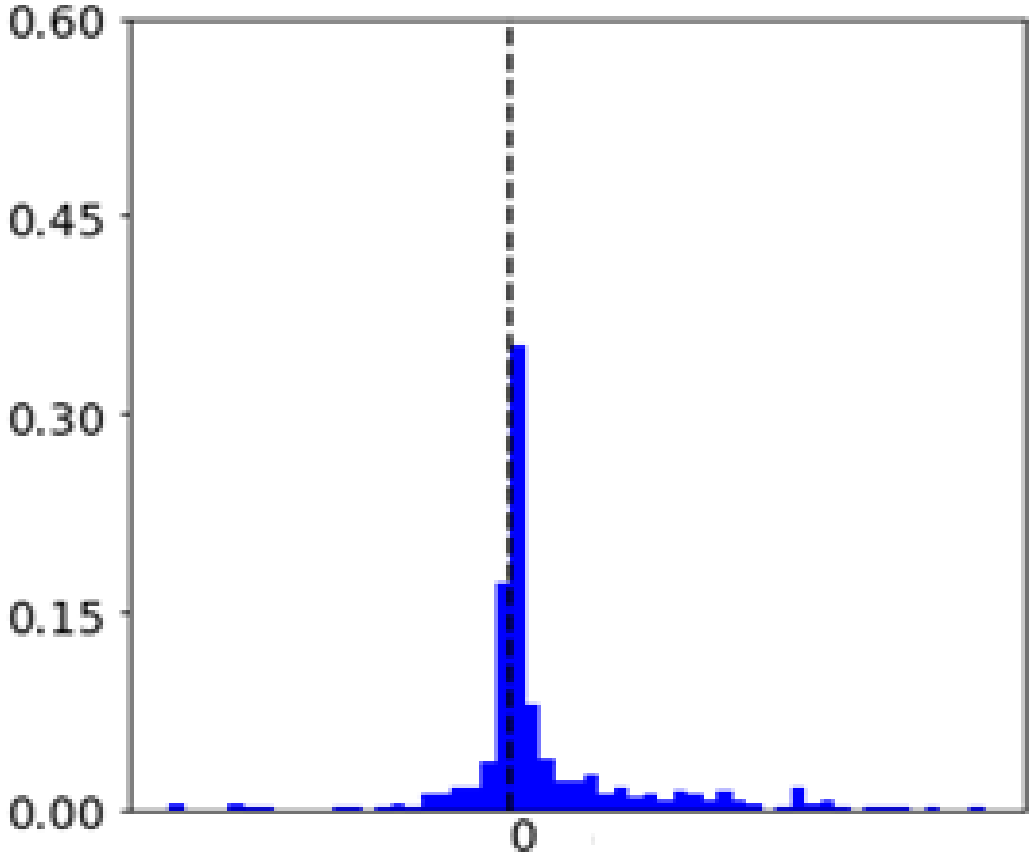
Results

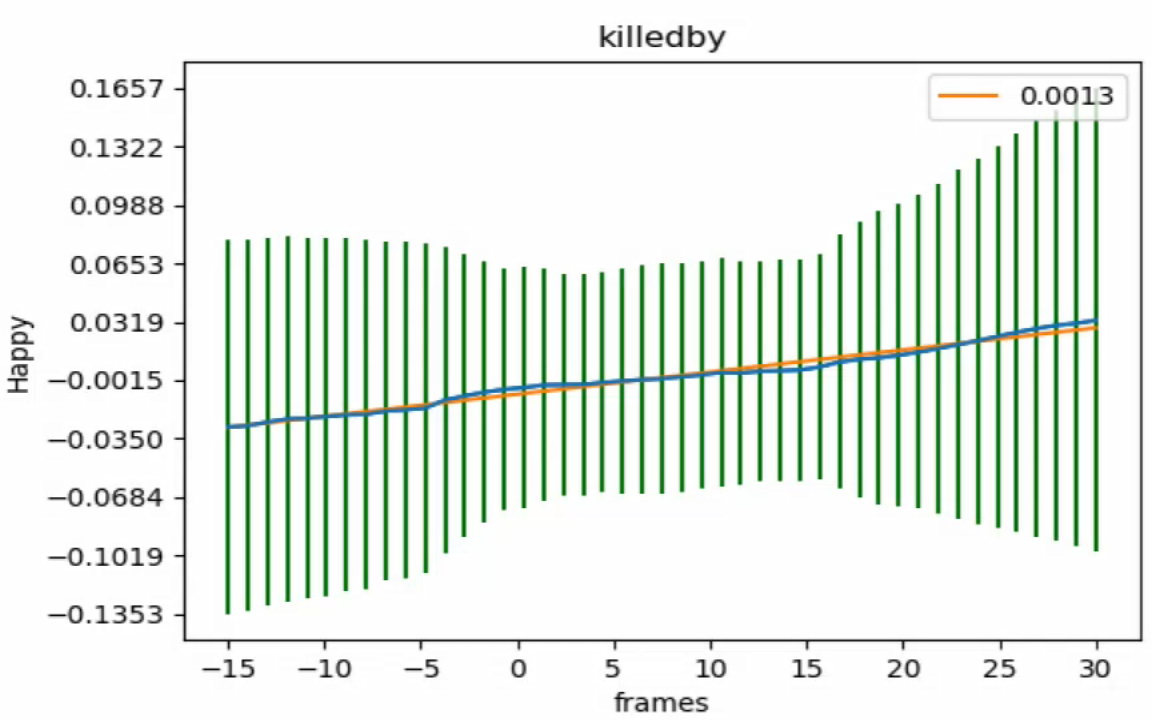
- Network applied to **Platformer Experience Dataset (PED)**
  - 58 players consisting of 28 males from Greece and Denmark, with ages ranging from 22 to 48 years.
  - Each player plays one or several games of **Infinite Mario Bros.**
  - Excel file for each video consisting of game events and their corresponding time stamp in the videos.
- **Exercise**: What makes players most happy when playing platformers?
  - Getting killed
  - Winning
  - Changing mode (Mario turning into Super Mario, or back)
  - Killing an enemy?

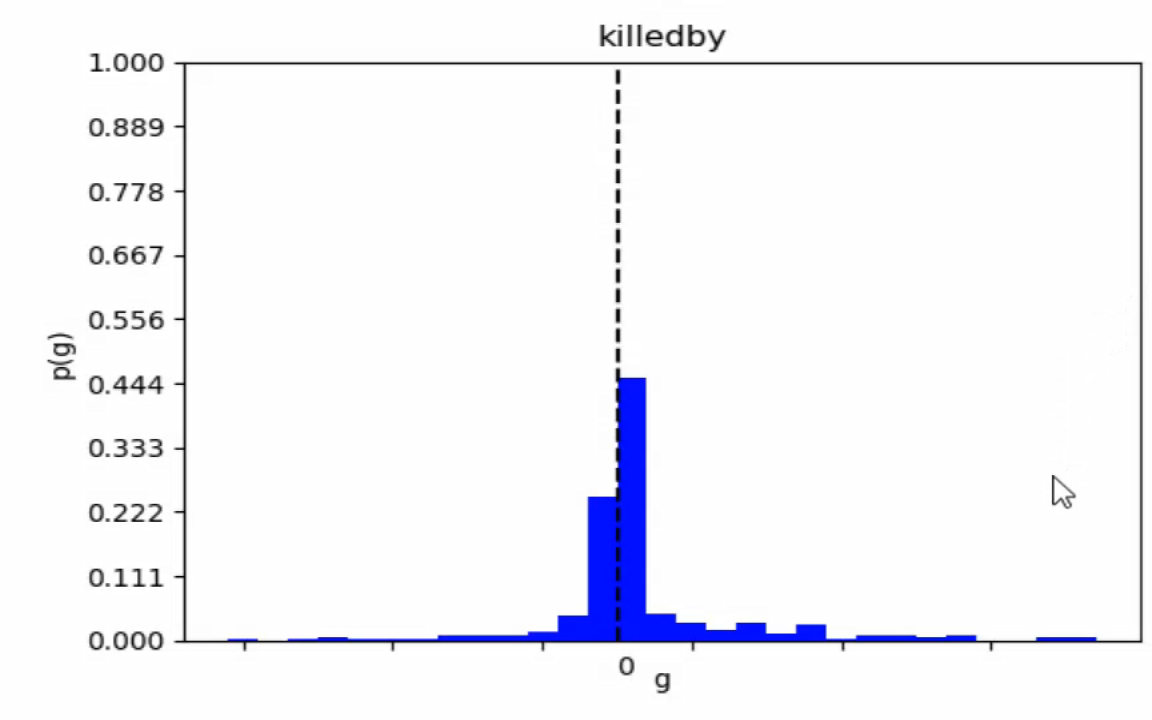
Affect gradient summary plots




killedby
0.0013
0.1810
0.1402
0.0994
0.0585
0.0177
−0.0231
−0.0639
−0.1047
−0.1455
−0.1863
Happy
−15
0
15
30
45
60
frames


0.60
0.45
0.30
0.15
0.00
0


killedby
0.0013
0.1657
0.1322
0.0988
0.0653
0.0319
−0.0015
−0.0350
−0.0684
−0.1019
−0.1353
Happy
−15
−10
−5
0
5
10
15
20
25
30
frames


killedby
1.000
0.889
0.778
0.667
0.556
0.444
0.333
0.222
0.111
0.000
p(g)
0
g

Getting Killed Makes Players Happy, According to a Neural Network

Shaghayegh Roohi, Jari Takatalo, J. Matias Kivikangas, Perttu Hämäläinen

SO MUCH DEATH!! | Super Meat Boy

1.9M views • 7 years ago

jacksepticeye

If you enjoyed the video, punch that LIKE button in the FA

Super Meat Boy Forever - Gameplay Wal
World)

524K views • 10 months ago

ZackScottGames

Thanks for watching my **Super Meat Boy** Forever Gamepl
played ...

Super Meat Boy (Almost) Full OST

920K views • 8 years ago

Gayla McGee

There is a bit at the end with just the image, this is due to
(Intro ...

I FINALLY Played Super Meat Boy!

296K views • 1 year ago

ryukahr

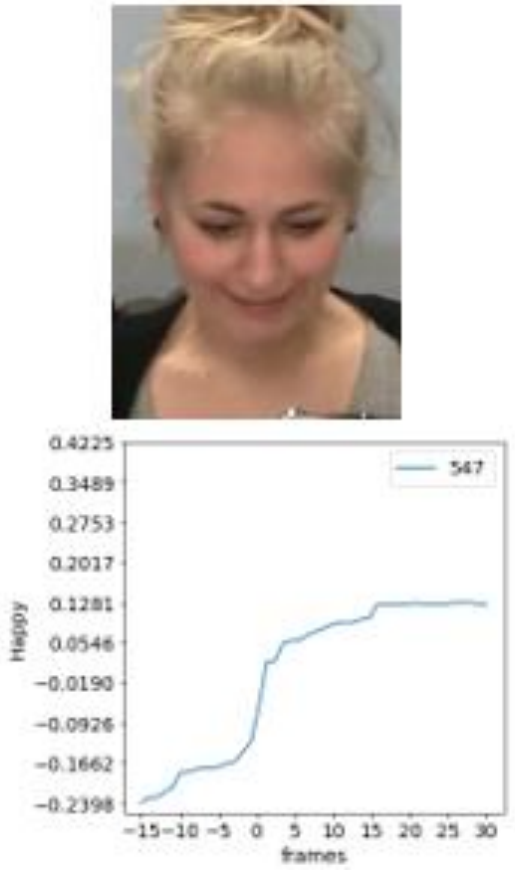
https://www.instagram.com/ryukahr/ It was fun looking for all of the things I missed while I was editing. TWITTER ...

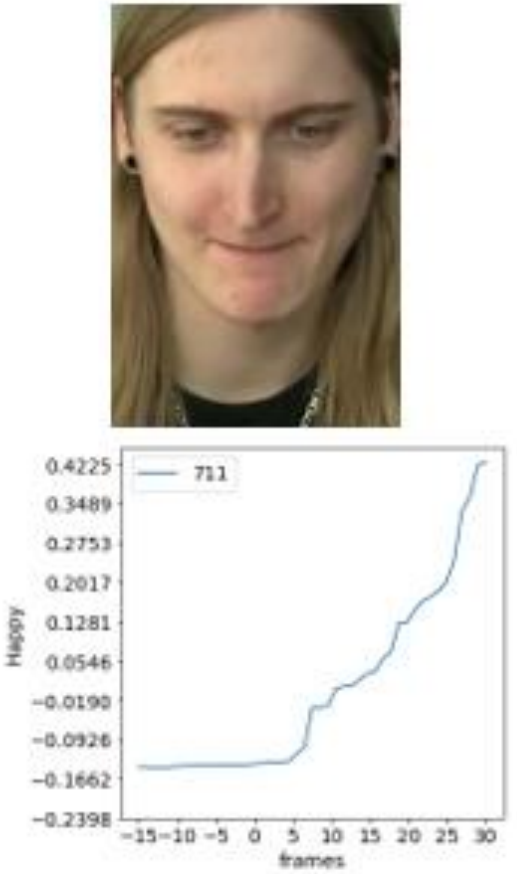
Is the goal to see how much blood I can drench the saws in?

Super Meat Boy Forever Review.
https://videochums.com/review/super-meat-boy-forever

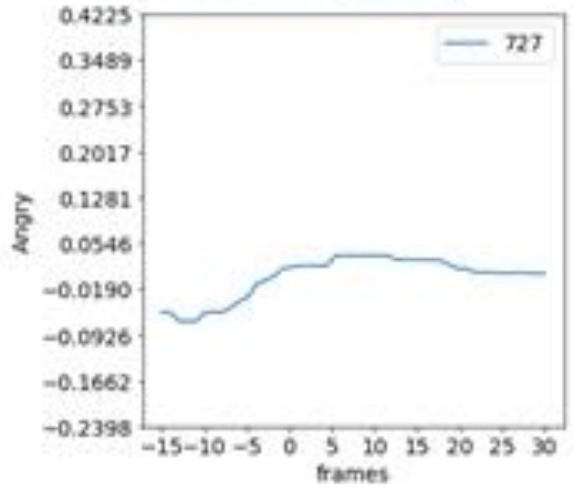
Results replicate previous psychophysiological studies

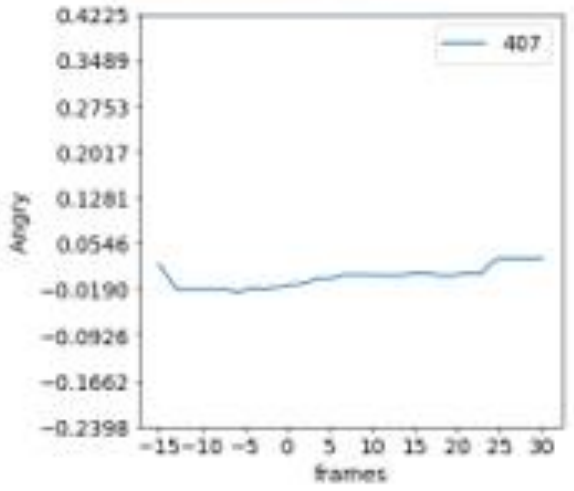
- Getting killed produces a smile

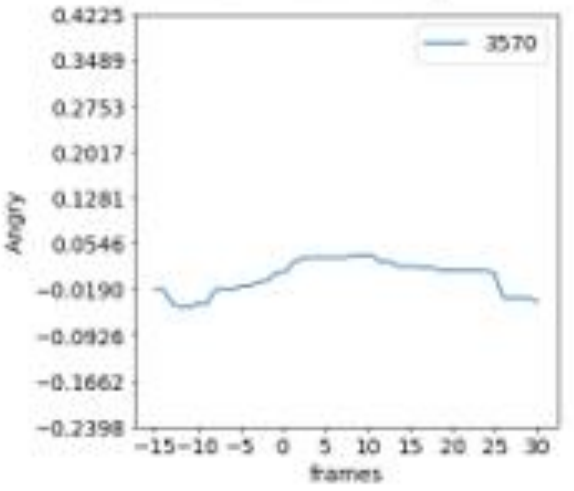




Killing enemies: concentrated frown interpreted as neutral or negative emotion







A Good Reason to Die: How Avatar Death and High Challenges Enable Positive Experiences

Serge Petralito[1], Florian Brühlmann[1], Glena Iten[1], Elisa D. Mekler[2] and Klaus Opwis[1]

[1]Center for Cognitive Psychology and Methodology, Department of Psychology, University of Basel
[2]HCI Games Group, Games Institute, University of Waterloo
{s.petralito, florian.bruehlmann, glena.iten, klaus.opwis}@unibas.ch, emekler@uwaterloo.ca

ABSTRACT

Appropriate challenges and challenge-skill balance are usually key to positive player experiences. However, some games such as the successful series *Dark Souls* are notorious for their excessive difficulty. Yet, there has been little empirical investigation of why players enjoy games they constantly struggle and fail with. We surveyed 95 participants right after the release of *Dark Souls III* about their experiences with the game, employing both open questions and different player experience measures. Players generally enjoyed challenging play sessions and mostly reported positive experiences, with achievement and learning moments strongly contributing to positive experiences. However, these factors themselves were enabled by negative events such as difficulties and avatar death. Our findings showcase that negative events bear a potential for forming positive and meaningful experiences, thus expanding previous knowledge about the role of challenge and failing in games. Moreover, the significance of hard-earned achievements extends present design conventions.

balance between challenge and skill. Hence, if challenge demands imposed by the game are too high or too low in regard to the player's skill level, playing the game leads to anxiety or boredom. The significance of an ideal challenge-skill balance is strongly emphasized in current research [3, 4, 13, 23, 33, 37], where adjustable and adaptive difficulty mechanics play an integral part in keeping this balance [8, 14, 35, 39]. Moreover, balance and accessibility represent two key notions of the *casual revolution*, a design trend towards making games more accessible by removing perceived barriers, penalties and frustrations and targeting much broader audiences than games used to over roughly a decade ago [20, 22]. In conclusion, challenge in current literature and modern game design has to a large extent been treated as a *Goldilocks factor*: The difficulty of a game should be neither too demanding nor too low in order to avoid negative experiences and frustrations.

In light of present design conventions, some exceptional games stand out, ignoring most of the conventional balancing efforts by implementing very high challenges and high consequential

Back-up: Self-Determination Theory

- Player's intrinsic motivation (predicting enjoyment, player persistence, etc.) based on satisfaction of 3 basic needs
- Autonomy, relatedness, and…
- **Competence**: need for challenge and feelings of effectance

Self-Determination Theory in HCI Games Research: Current Uses and Open Questions

April Tyack[1,2], **Elisa D. Mekler**[2]
[1]Queensland University of Technology (QUT), Brisbane, Australia
[2]Aalto University, Espoo, Finland
{firstname.lastname}@aalto.fi

ABSTRACT
Self-Determination Theory (SDT), a major psychological theory of human motivation, has become increasingly popular in Human-Computer Interaction (HCI) research on games and play. However, it remains unclear how SDT has advanced HCI games research, or how HCI games scholars engage with the theory. We reviewed 110 CHI and CHI PLAY papers that cited SDT to gain a better understanding of the ways the theory has contributed to HCI games research. We find that SDT, and in particular, the concepts of need satisfaction and intrinsic motivation, have been widely applied to analyse the player experience and inform game design. Despite the popularity of SDT-based measures, however, prominent core concepts and mini-theories are rarely considered explicitly, and few papers engage with SDT beyond descriptive accounts. We highlight conceptual gaps at the intersection of SDT and HCI games research, and identify opportunities for SDT propositions, concepts, and measures to more productively inform future work.

Author Keywords
Games; Gamification; Motivation; Play; Player Experience; Self-Determination Theory; Theory

CCS Concepts
•**Human-centered computing** → **HCI theory, concepts and models;** Empirical studies in HCI; •**Applied computing** → *Computer games;*

INTRODUCTION
One aim of games and play research in Human-Computer Interaction (HCI) – hereafter abbreviated to HCI games research – is to understand what constitutes engaging player-computer interaction [117]. These insights may in turn be applied to design more appealing games and playful interactions, evaluate qualities of the player experience, and create interactive systems that motivate people to engage with purposes beyond entertainment (e.g., serious games, gamification). Theories and concepts from motivational psychology have proven particularly popular with HCI scholars to describe and analyse games [30, 99]. The notion of *flow* [128], for instance, has been influential in studying the player experience [79, 124, 177] and modelling optimally challenging games [40, 113]. Another theory that has proven influential is Self-Determination Theory (SDT), a major psychological theory of human motivation [48, 163] that has been successfully applied to study motivational processes in a variety of domains and contexts (e.g., academic, work, relationships). SDT has been used to study the motivational appeal of games [160, 166], inform gameful design [60, 149, 187], analyse the player experience [93, 150], and applied within the games industry for evaluation and testing [7, 85, 189]. In fact, the original papers on SDT and games by Ryan, Rigby and Przybylski [144, 166] have been cited over 3000 times on Google Scholar.

While these numbers attest to the popularity of (citing) SDT in games research, they say little about *how* SDT has contributed to HCI games research, nor the ways in which HCI games scholars have applied and engaged with the theory. Some of the purported benefits of applying (psychological) theories to HCI include establishing a common understanding and terminology around specific phenomena, formulating predictions concerning these phenomena under common and novel circumstances, as well as generating original hypotheses and design implications [13, 132, 153]. However, the extent to which SDT has informed HCI games research remains unclear. Concerns have also been raised around the misrepresentation of external literature in exertion games research [120], and with respect to SDT in gamification research [116, 169]. Not only does this risk the proliferation of misunderstandings and lack of clarity regarding SDT-based concepts – it may also give rise to invalid research findings around the motivational appeal of games, ineffective design implications, or even negative effects on player wellbeing.

Following endeavours on the use of theory in HCI [38, 121, 153, 195], we present findings from a systematic literature review encompassing 110 CHI and CHI PLAY papers that cite SDT in the context of games, play, and game-adjacent systems. We take stock of how and why SDT and its various concepts (e.g., intrinsic motivation, need satisfaction) have been applied to HCI games research. Our contribution is threefold: first,

Questions?

Aalto University
School of Science
and Technology

Recognizing Emotional Expression in Game Streams

Shaghayegh Roohi - shaghayegh.Roohi@aalto.fi
Elisa D. Mekler - elisa.mekler@aalto.fi
Mikke Tavast - mikke.tavast@aalto.fi
Tatu Blomqvist - tatu.blomqvist@aalto.fi
Perttu Hämäläinen - perttu.hamalainen@aalto.fi

Motivation

- Importance of emotions in games
  - Yannakakis and Paiva [1] argued that “one cannot dissociate games from emotions” (p. 459).
  - An active research area: Emotional attachment to game characters, emotional challenge, grief & other negative emotions can produce positive experiences…
- An underexplored data trove: game streaming videos
- Streamers narrate what they do, show emotion (no “gamer face”), both game and face in the same video

[1] Georgios N Yannakakis and Ana Paiva. 2014. Emotion in games. Handbook on affective computing (2014), 459–471.
[2] 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.
[3] 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.
[4] 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.

An example of game stream videos

Contribution

- A dataset of human-annotated emotional expression in game streams
  - 17 videos, 11 hours, and 2015 emotional events, on average one annotated event for each 40 seconds of video
- A multimodal neural network to mimic human annotations
  - Facial expression (7 classes of emotion)
  - Video transcript sentiment analysis (positivity of speech)
  - Voice emotion analysis (7 classes of emotion)
  - Voice features (e.g., loudness and pitch)

Dataset preparation

- Streams of the games Unravel[1] and its sequel, Unravel Two [2]
  - Puzzle platformers games
- Reason of game selection
  - Recently released and readily featured on several YouTube channels.
  - Linear level design, where all players experience game events in the same sequence.
  - Both Unravel and Unravel Two were praised for being emotionally engaging [3, 4]

[1] Coldwood Interactive. 2016. Unravel. Game [PC, PlayStation 4, Xbox One]. (February 2016). Electronic Arts.
[2] Coldwood Interactive. 2018. Unravel Two. Game [PC, PlayStation 4, Xbox One]. (June 2018). Electronic Arts.
[3] 2019a. Unravel for PC Reviews - Metacritic. (April 2019). https://www.metacritic.com/game/pc/unravel Retrieved April 5 2019.
[4] 2019b. Unravel Two for PC Reviews - Metacritic. (April 2019). https://www.metacritic.com/game/pc/unravel-two Retrieved April 5 2019.

Dataset preparation

- Criteria for stream selection
  1. The streamer’s face had to be visible throughout the video
  2. The streamer provided commentary in English
  3. Only one person was playing and present during the stream
  4. Available subtitle transcripts from automatic captioning
- 17 videos by 9 different streamers (2 women, 7 men)
- Each video has been annotated by two persons into 13 classes of emotion like amusement, frustration
- Each emotional event has been labeled as top5 event if it is among the top5 events of the stream

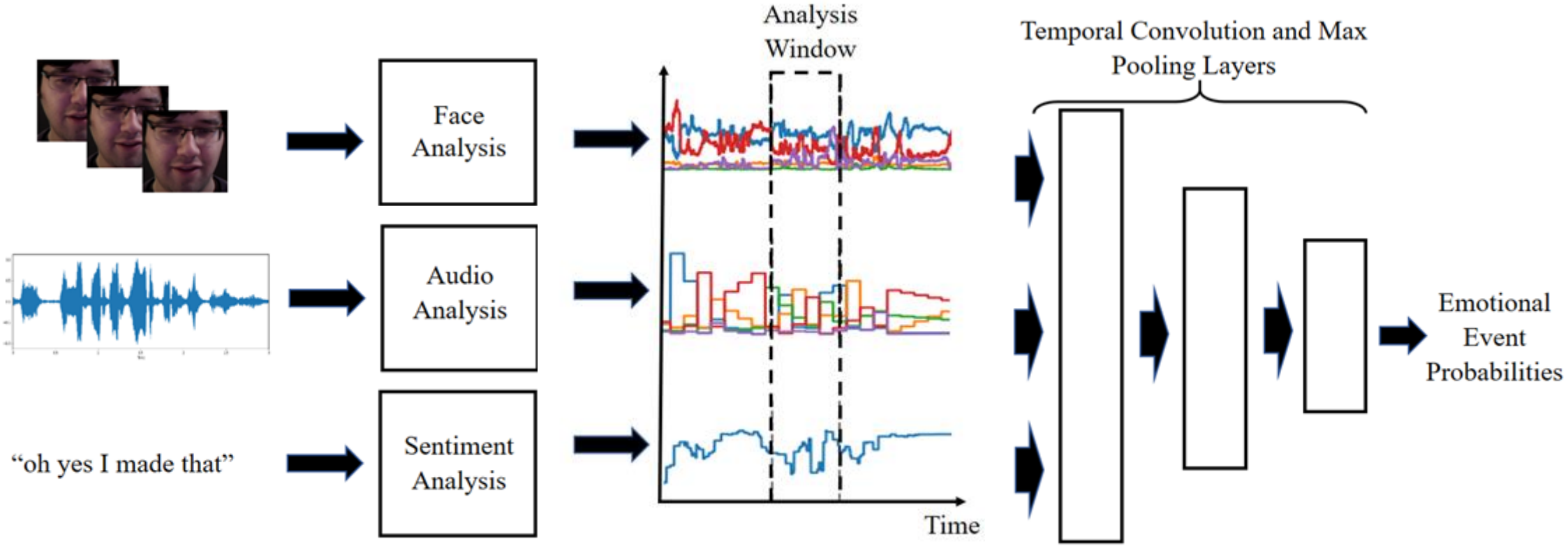
Human Inter-rater agreement

- 2 classes (no event or event)
- 2 classes/top5 (top 5 events or not top 5 events)
- 4 classes (no event, pleasant event, unpleasant event, and neutral event)
- 14 classes (no event and the full set of 13 codes)

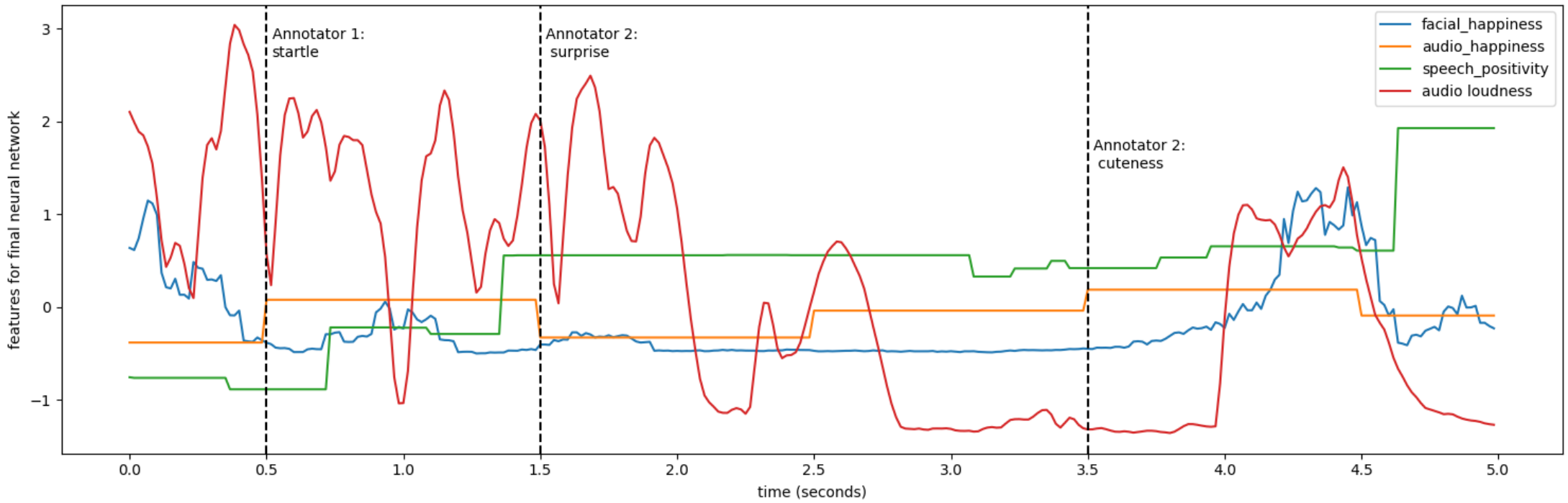
Window length	Inter-rater agreement			
	2-class	2-class-top events	4-class	14-class
1	59.6	53.4	34.9	18.8
2	64.3	56.8	39.3	24.0
3	67.0	59.1	41.8	27.3
4	68.3	61.0	43.1	29.4
5	68.7	60.3	43.7	30.8

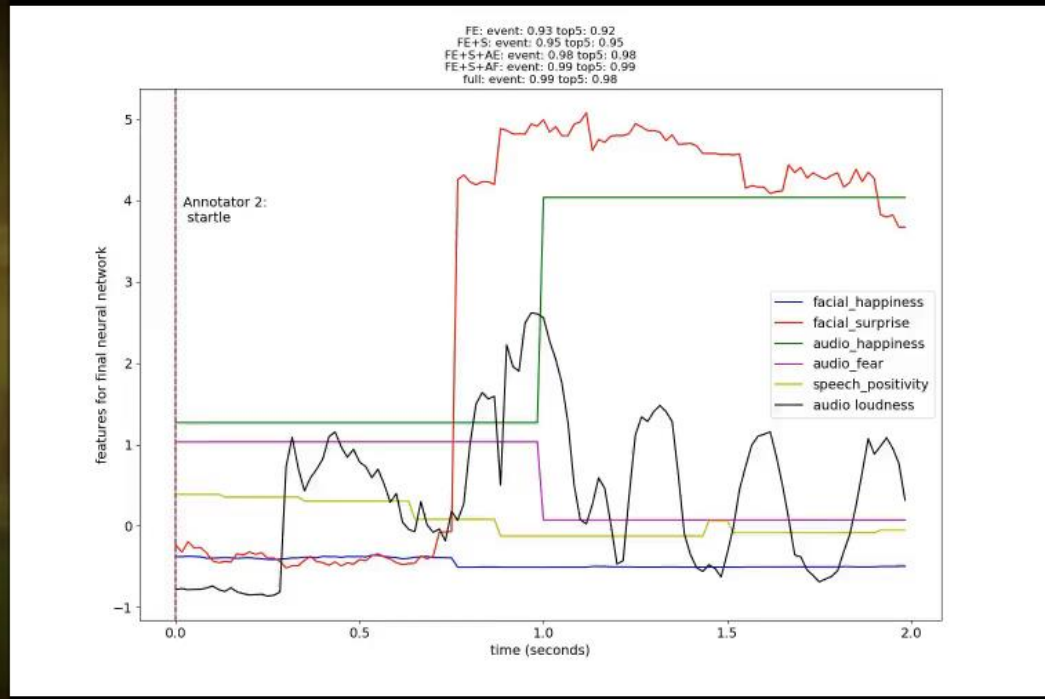
Table 1. Inter-rater agreement and congestion with respect to different window lengths and levels of granularity

Automatic emotional event detection



An example of the multimodal input signals




FE: event: 0.93 top5: 0.92
FE+S: event: 0.95 top5: 0.95
FE+S+AE: event: 0.98 top5: 0.98
FE+S+AF: event: 0.99 top5: 0.99
full: event: 0.99 top5: 0.98
Annotator 2:
startle
facial_happiness
facial_surprise
audio_happiness
audio_fear
speech_positivity
audio loudness
features for final neural network
time (seconds)
5
4
3
2
1
0
−1
0.0
0.5
1.0
1.5
2.0

Results: Scoring

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

- F1 score:

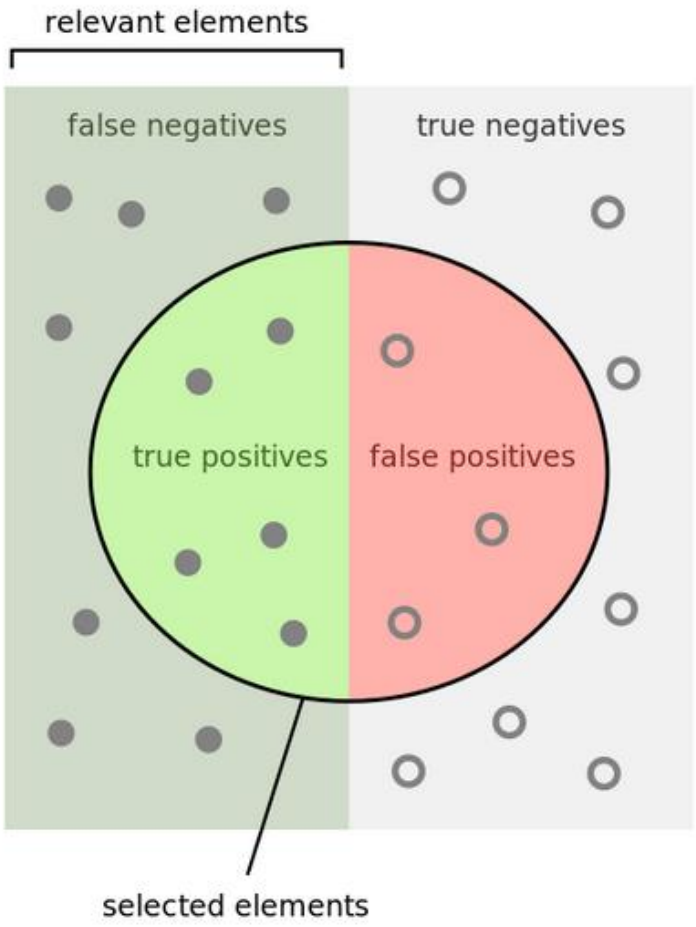
$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}.$$

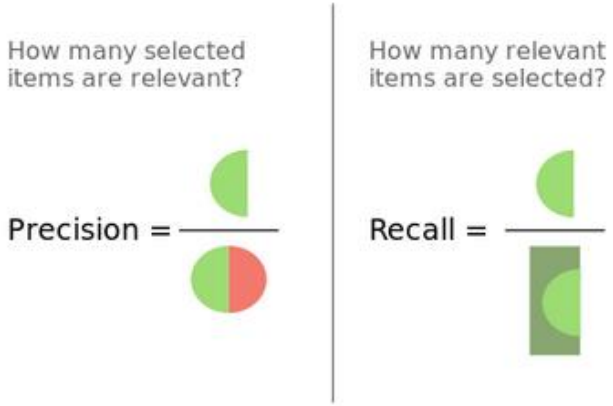
precision or positive predictive value (PPV)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}$$

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR}$$





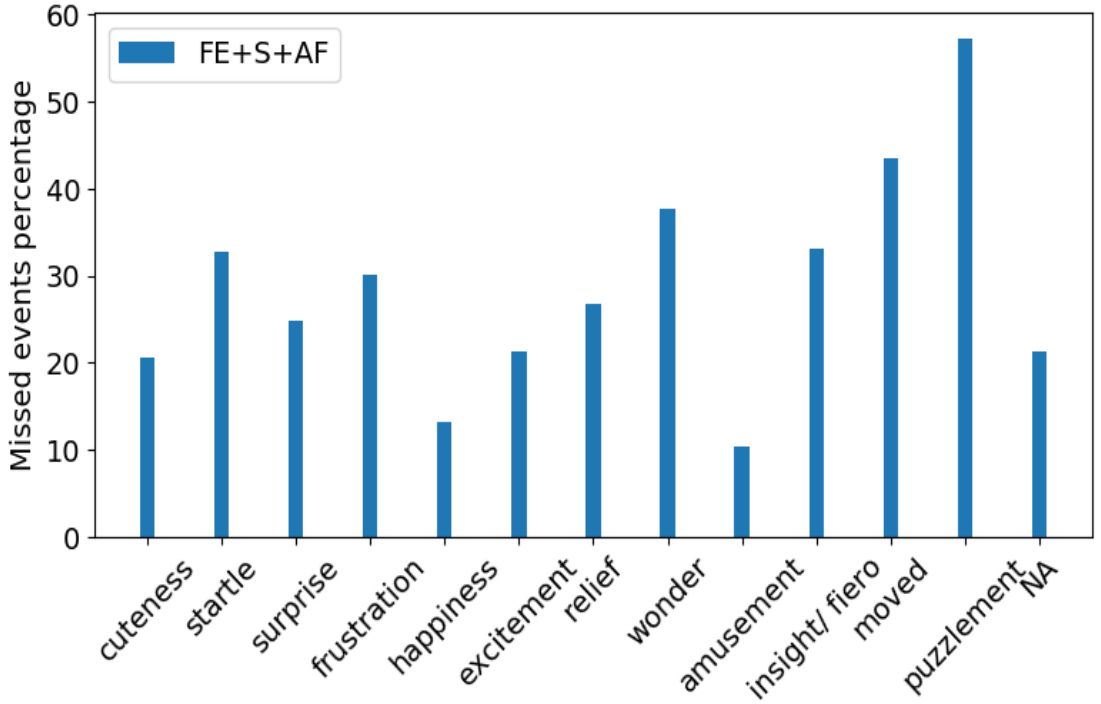
Automatic emotion detection results

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.

Percentage of missed events

- Most accurate at recognizing positive emotional expressions
- False negatives were more prevalent for subtle emotional expressions
  - may be due to lower intensity expressions
- High amount of false negatives for Frustration and startle might be due to being accompanied by immediate positive expressions (e.g., smiling)



Future work

- Automatic annotation of gameplay events as a complementary modality (can train a neural network to recognize player death etc)
- Investigate effect of gender, age or cultural differences in large datasets of streamers
- Test our approach with playtest videos recorded without game audio that can interfere with audio expression analysis

Summary

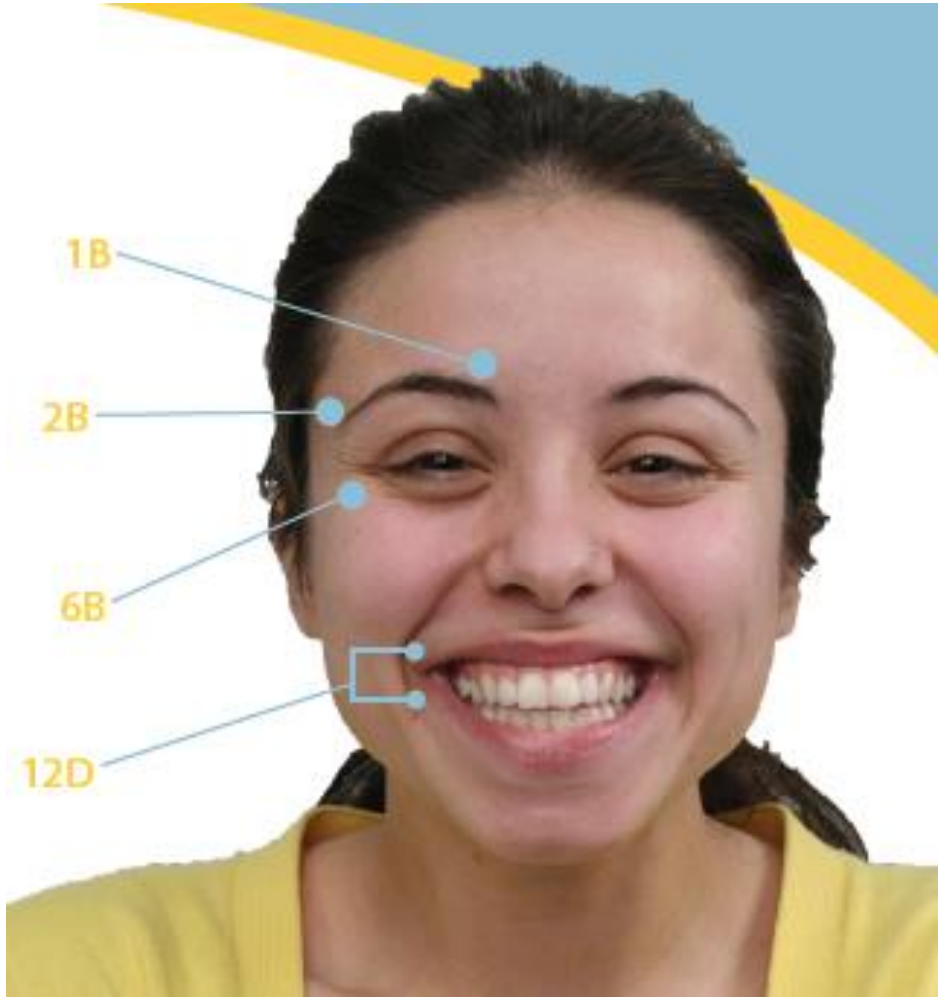
- We have presented a new dataset of emotional events
- We have presented automated detection of emotionally salient events in game stream videos
- Identifying and classifying emotional events is a task that is hard for both humans and artificial neural networks.
- Using limited number of classes, we can achieve promising results
  - Only detecting the events yields a decent automatic detection accuracy of 70.7% which is on par with human inter-rater agreement of 68.7%
  - Applications in video highlights detection or pre-selecting videos for further analysis

Questions?

Practicalities:

- The CHI PLAY 2019 paper describes datasets and network architectures for processing the audio, video, and transcript data: https://dl.acm.org/citation.cfm?id=3347197
- Currently recommended facial expression analysis tool that we found after writing the paper: OpenFace. Simple command-line tools, no need to train any neural networks oneself. https://github.com/TadasBaltrusaitis/OpenFace/wiki

Facial Action Coding System (FACS)



Afternoon Exercise

Exercise

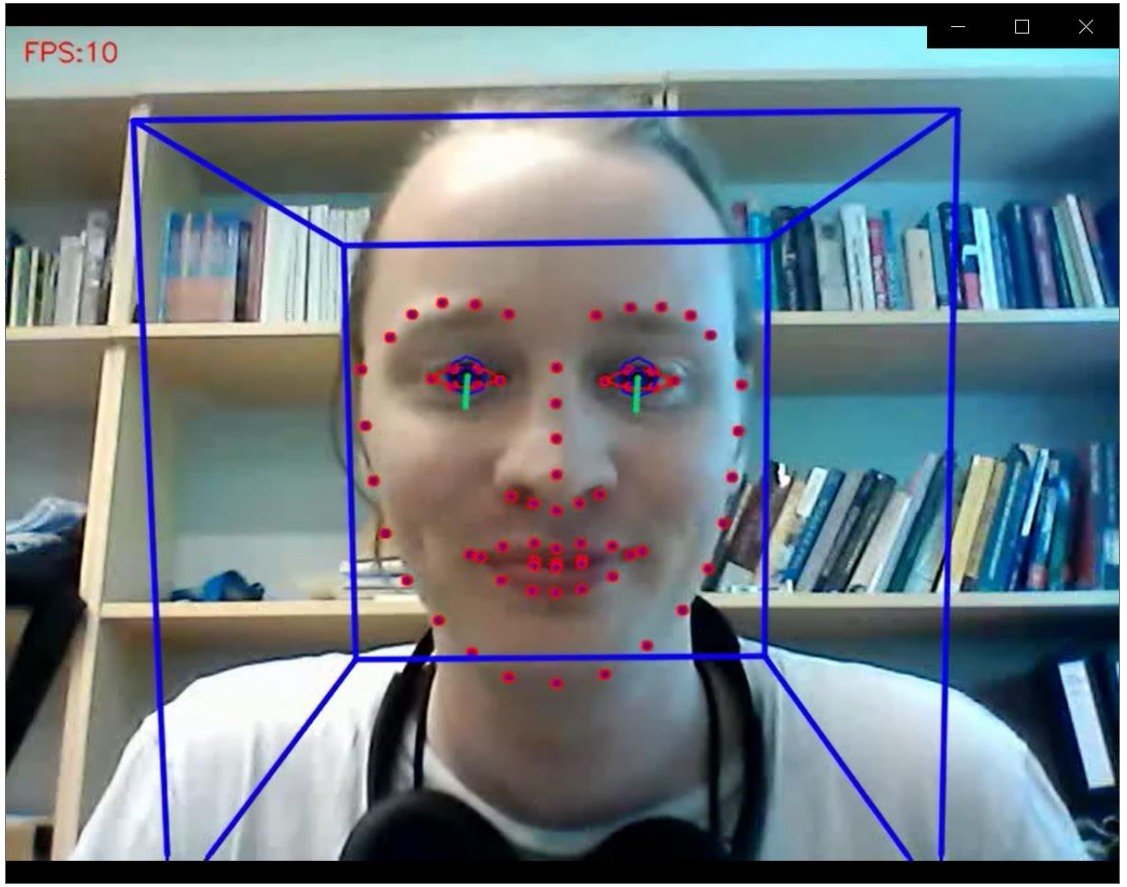
- Teams of 3-5, at least 1 game-ready Laptop per team
  - Split into experimenter(s) and participant(s)
- Experimenters: design a facial expression experiment
  - Pick one or two games (your own, Steam, Twitch)
  - Formulate closed or open-ended research question
  - Put forward your hypothesis: what do you expect to find?
- Participants: play game, have yourself recorded
- All:
  - Analyse result with OpenFace (command line or GUI!)
  - Report study in slidedeck with at most 3 slides: (1) game (video?), question, hypothesis; (2) results; (3) discussion of your results.
  - Select a presenter and **present today 15.00, back here.**

Tip:
Try out OpenFace before posing your research question! What data can you collect?

OpenFace 2.2.0: a facial behavior analysis toolkit

https://github.com/TadasBaltrusaitis/OpenFace/wiki

- Openface is an easy to use opensource toolkit that detects facial landmarks, head pose, eye-gaze direction and facial action units
- Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019



Action unit estimation

- Openface can estimate facial movements based on Facial Action Coding System (FACS)
- In FACS, facial movements are coded as different action units (AU): https://imotions.com/blog/facial-action-coding-system/
- Openface estimates the presence and the intensity (scale from 0 to 5) of different AUs in each frame of the video
- For the accuracy of AU estimates, see Baltrusaitis, Zadeh, Lim, & Morency (2018)

How to use, step by step (Windows Powershell)

- Openface is operated trough command line interface: e.g. with PowerShell (Windows) or xterm (Unix)
- 1. Install: https://github.com/TadasBaltrusaitis/OpenFace/wiki/Windows-Installation
- 2. Open Windows Powershell
- 3. Change the directory to the openface folder, for example:
  - cd C:
  - cd \...\...\...\OpenFace_2.0.5_win_x64\

Useful command line arguments

- 4: Execute a command, examples:
  - Extract features (CSV-file): .\FeatureExtraction.exe -f "C:\...\...\filename.avi”
  - Extract AU estimates (CSV-file): \FeatureExtraction.exe -aus –f "C:\...\...\...\filename.avi”
  - Visualize the data: .\FeatureExtraction.exe –verbose "C:\...\...\filename.avi”
- 5. Analyze the data ☺
- Also possible to use GUI with the argument ./OpenFaceOffline.exe
- List of all possible arguments: https://github.com/TadasBaltrusaitis/OpenFace/wiki/Command-line-arguments

Administrator: Windows PowerShell

```
PS C:\Openface\OpenFace_2.0.5_win_x64> .\FeatureExtraction.exe -f "C:\my videos\video.avi" -verbose
Reading the landmark detector/tracker from: model/main_ceclm_general.txt
Reading the landmark detector module from: model\cen_general.txt
Reading the PDM module from: model\pdms/In-the-wild_aligned_PDM_68.txt....Done
Reading the Triangulations module from: model\tris_68.txt....Done
Reading the intensity CEN patch experts from: model\patch_experts/cen_patches_0.25_of.dat....Done
Reading the intensity CEN patch experts from: model\patch_experts/cen_patches_0.35_of.dat....Done
Reading the intensity CEN patch experts from: model\patch_experts/cen_patches_0.50_of.dat....Done
Reading the intensity CEN patch experts from: model\patch_experts/cen_patches_1.00_of.dat....Done
Reading part based module....left_eye_28
Reading the landmark detector/tracker from: model\model_eye/main_clnf_synth_left.txt
Reading the landmark detector module from: model\model_eye\clnf_left_synth.txt
Reading the PDM module from: model\model_eye\pdms/pdm_28_l_eye_3D_closed.txt....Done
Reading the intensity CCNF patch experts from: model\model_eye\patch_experts/left_ccnf_patches_1.00_synth_lid_.txt....Done
Reading the intensity CCNF patch experts from: model\model_eye\patch_experts/left_ccnf_patches_1.50_synth_lid_.txt....Done
Done
Reading part based module....right_eye_28
Reading the landmark detector/tracker from: model\model_eye/main_clnf_synth_right.txt
Reading the landmark detector module from: model\model_eye\clnf_right_synth.txt
Reading the PDM module from: model\model_eye\pdms/pdm_28_eye_3D_closed.txt....Done
Reading the intensity CCNF patch experts from: model\model_eye\patch_experts/ccnf_patches_1.00_synth_lid_.txt....Done
Reading the intensity CCNF patch experts from: model\model_eye\patch_experts/ccnf_patches_1.50_synth_lid_.txt....Done
Done
Reading the landmark validation module....Done
Reading the AU analysis module from: AU_predictors/main_dynamic_svms.txt
Reading the AU predictors from: AU_predictors\AU_all_best.txt... Done
Reading the PDM from: AU_predictors\In-the-wild_aligned_PDM_68.txt... Done
Reading the triangulation from:AU_predictors\tris_68_full.txt... Done
Attempting to read from file: C:\my videos\video.avi
Device or file opened
Starting tracking
Reading the MTCNN face detector from: model/mtcnn_detector/MTCNN_detector.txt
Reading the PNet module from: model/mtcnn_detector\PNet.dat
Reading the RNet module from: model/mtcnn_detector\RNet.dat
Reading the ONet module from: model/mtcnn_detector\ONet.dat
0% 10% 20% 30% 40% 50% 60% 70% Closing output recorder
Closing input reader
Closed successfully
Postprocessing the Action Unit predictions
PS C:\Openface\OpenFace_2.0.5_win_x64> .\FeatureExtraction.exe -f "C:\my videos\video.avi" -verbose
```

Add-in Express

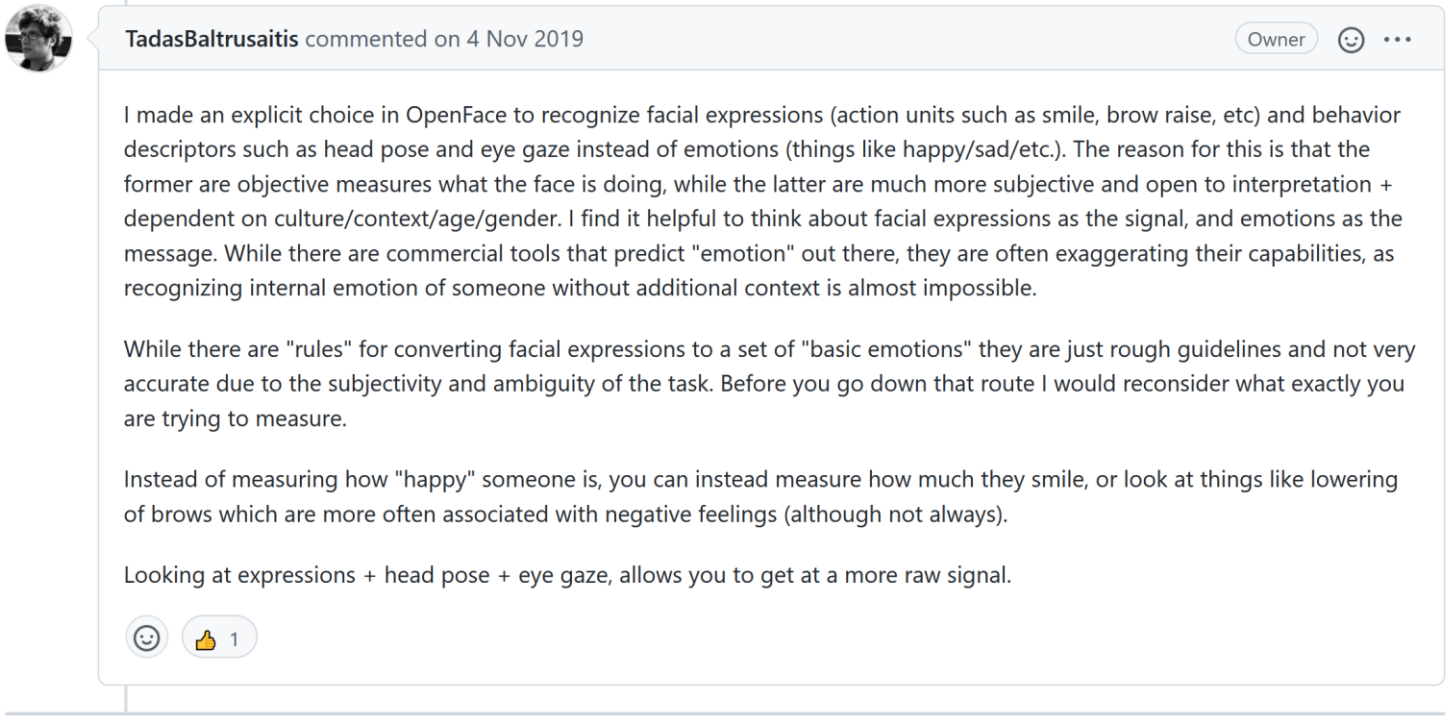
GPowerWin_3.1.9.2

18.42
11.6.2019

From Action Units to Emotions

Emotion	Action units
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+17
Contempt	R12A+R14A

You will (i) either have work with action units only, or (ii) map from action units to emotions (thus: better focus on 1 / 2 emotions only)



TadasBaltrusaitis commented on 4 Nov 2019 — Owner

I made an explicit choice in OpenFace to recognize facial expressions (action units such as smile, brow raise, etc) and behavior descriptors such as head pose and eye gaze instead of emotions (things like happy/sad/etc.). The reason for this is that the former are objective measures what the face is doing, while the latter are much more subjective and open to interpretation + dependent on culture/context/age/gender. I find it helpful to think about facial expressions as the signal, and emotions as the message. While there are commercial tools that predict "emotion" out there, they are often exaggerating their capabilities, as recognizing internal emotion of someone without additional context is almost impossible.

While there are "rules" for converting facial expressions to a set of "basic emotions" they are just rough guidelines and not very accurate due to the subjectivity and ambiguity of the task. Before you go down that route I would reconsider what exactly you are trying to measure.

Instead of measuring how "happy" someone is, you can instead measure how much they smile, or look at things like lowering of brows which are more often associated with negative feelings (although not always).

Looking at expressions + head pose + eye gaze, allows you to get at a more raw signal.

1

Webcam Video Recording

- E.g. “Camera App” in Windows 10
  - https://www.digitalcitizen.life/how-use-camera-app-windows-10-your-webcam/
- Alternatively: VLC Player
  - https://www.videolan.org/vlc/
  - https://www.vlchelp.com/how-to-record-webcam-video-using-vlc-media-player/

Recap

- Categorical emotion classification or positive/negative sentiment analysis is fairly straightforward from video, voice, and speech transcripts
- Not very nuanced, but still useful data!
- Limitation: trained on human data, which can be noisy and with low intercoder agreement.
- OpenFace gives the most nuanced facial expression data (FACS activations)
- Relation of facial expression and emotions is complex