Streaming Behaviour: Live Streaming as a Paradigm for Multi-view Analysis of Emotional and Social Signals

Mihalis A. Nicolaou
Department of Computing
Goldsmiths, University of London
London, United Kingdom

Charles Ringer
Department of Computing
Goldsmiths, University of London
London, United Kingdom

ABSTRACT

We argue that live video game streams constitute an invaluable paradigm towards building multi-view, data-driven models of human behaviour. The interactive setting under which a stream operates is enriched with social signals, conveyed between streamers and viewers via facial expressions, body movement, vocal cues, as well as written language. We consider the data sources involved in a typical broadcast (e.g., camera stream, game footage, text) as data-views that carry inherent correlations, since they all describe events occurring during a stream. We argue that this unique interactive setting facilitates the *joint*, multi-view analysis of various heterogeneous data sources involved, in a coherent and self-contained manner. We elaborate on the emergence of social signals in this setting, while discussing close links and potential research directions related to areas such as *affective computing*, *machine learning*, *computer vision* and *intelligent game design.*

1 INTRODUCTION

Video game live streaming has become a dominant form of entertainment over the past years, with emerging communities of millions of broadcasters and viewers on popular platforms such as TWITCH.TV [5]. This paper is motivated by two relevant observations, namely (i) the inherently interactive nature of streaming, where streamers may interact with each other as well as their viewers, and (ii) the rich, heterogeneous data that accompanies a broadcast (e.g., game and webcam footage, speech, natural language). We argue that the social, interactive setting that streamers find themselves in leads to the emergence of a multiset of complex emotional and social signals (e.g., during co-operative play), that can be studied in the context of *affective computing*, where research in the past decade has focused on the machine analysis of social and emotional signals based on *visual* and *auditory* cues. The 'streamer's paradigm' provides a rich source for vast amounts of heterogeneous data that encode social signals, as conveyed by facial expressions, speech emotion, as well as viewer sentiment via live chat. As we discuss in what follows, this setting gives rise to a set of research problems that strongly intersect with research undertaken in areas such as affective computing, machine learning and computer vision, audio signal processing and natural language processing, as well as intelligent game design.

2 RELATED WORK

Most related to the live stream setting are databases collected towards player experience modelling [15]. For instance, the Platformer Experience Dataset (PED) [3] employs games as a means to elicit affect, and contains audio-visual recordings of players along with in-game data, context and self-reported experience rankings. Main variations with respect to the game stream setting include (i) the *social interaction* element, where players and viewers interact and therefore are more likely to communicate social and emotional cues, and (ii) the uncontrolled settings under which streams take place. Other works analysing streamer behaviour include [10], where webcam and physiological data have been used for the analysis. Nevertheless, this work focuses on building a tool that utilizes existing SDKs for facial analysis such as Affdex [8], while wristbands are used for extracting physiological data - thus requiring access to the streamer. Other related works include analysis of affect and behaviour in co-operative and single player games, c.f. [1], where physiological and visual data from eight subjects is utilized. Finally, we refer the reader to recent articles such as [12, 13] for a review on related tasks such as audio-visual analysis of human behaviour and facial expression analysis, while the interaction between machine learning and intelligent game design is further explored in [16].

3 STREAMING & MULTI-VIEW DATA

In this section, we discuss data sources emerging in a live stream setting, while considering social interactions taking place. During a typical game stream, the streamer broadcasts webcam video, usually capturing the streamer's face and in some cases body, along with audio and the game footage. Social and behavioural cues are communicated to viewers and potentially other streamers with facial expressions, body gestures, as well as via written and spoken language. Viewers communicate via chat, as well as by raising various stream events (e.g., subscriptions). We expect that the interactive setting assumed in a typical game stream will give rise to more rich behavioural and emotional content. In settings where streamer-to-streamer interaction arises (e.g., co-operative play), the data sources available extend to cover the broadcast of other players participating. By considering both streams (including facial expressions and audio of both players), we can potentially better analyse more complex social behaviours such as *conflict*, *agreement*, and *mimicry*. In Fig. 1, we illustrate the discussed game stream concept from a data-centric perspective. The audio-visual[1] nature of the stream points to research in machine learning, computer vision, and acoustic signal processing, while data in the form of written language to areas such as sentiment analysis and natural language processing. By considering the *uncontrolled* settings under which broadcasts are taking place, the analysis of game streams becomes a very challenging and rich data paradigm.

In Fig. 2, we show camera stills from data collected during a typical game stream of a popular game. In more detail, in the first and second rows, frames indicate the variability in terms of facial

[1]We note that fusing audio-visual data is in itself a challenging research direction [4].

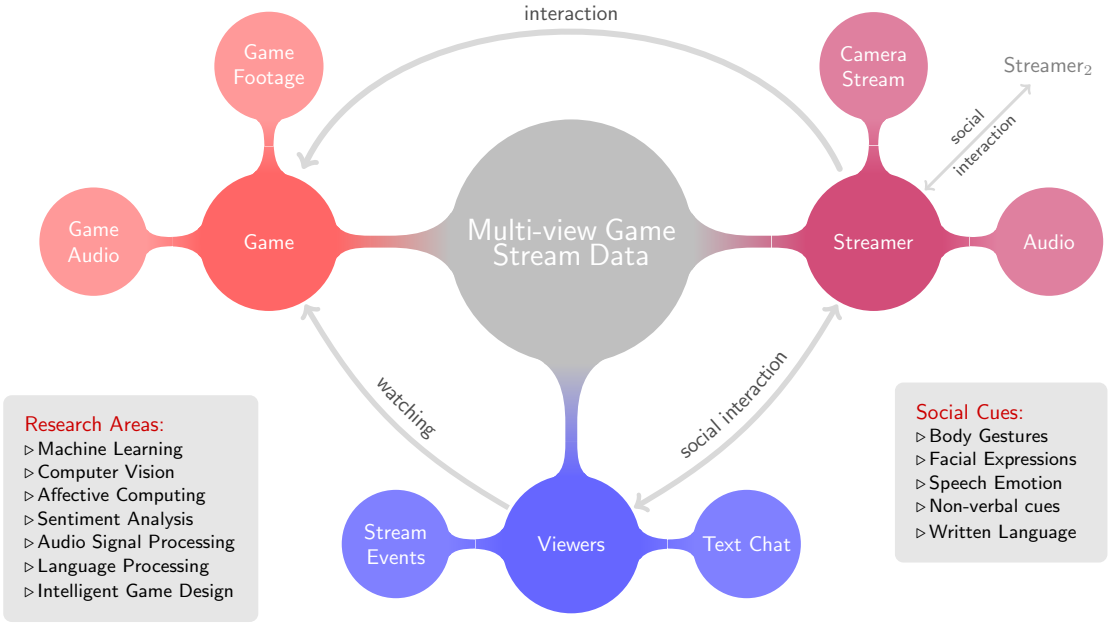

Figure 1: Illustration of main concepts discussed in the paper, including social interactions, social and emotional signals conveyed, as well as related research areas. The streamer broadcast includes a video stream capturing face and/or body, along with audio recorded from the streamer's microphone. This is accompanied by the footage of the game, including in-game audio. The streamer is interacting with viewers by video, voice, and text. Social signals can be extracted by analysing facial expressions, body gestures, speech emotion and non-verbal cues. The viewers interact with streamers and other viewers via text and stream-specific events, where analysis of the former is in the realms of sentiment and natural language processing. A frequent setting includes the streamer interacting with other streamers via webcam video and voice (e.g., during a co-op game). This allows for modelling complex dyadic interactions via voice and facial expression analysis using data arising from both broadcasts.

Figure 2: Camera stills from a typical game stream, including various facial expressions, body gestures, head poses, as well as challenging frames with face occlusions.

expressions and gestures respectively, as expressed by the streamer while communicating with viewers and other streamers. The third and fourth rows contain stills capturing head pose variability, as well as randomly selected stills where parts of the face are occluded by other body parts. This deems the computer vision analysis of facial data even more challenging, and is also on par with the recent developments in face analysis, where data captured under unconstrained settings ('in-the-wild') are commonly considered as challenging benchmarks [6, 7].

4 RESULTS

While stream data is in abundance, providing annotations on which machine learning models can be trained is a tedious and laborious task, that requires some form of human supervision and is particularly expensive in case per-frame labels are to be provided. It is therefore reasonable to firstly approach this problem in the context of unsupervised learning. In our recent work [9], we propose a deep unsupervised novelty-detection approach for detecting highlights in game streams[2], by utilizing multi-view data including streamer facial expressions, audio, and game footage.

The method consists of two convolutional autoencoder networks encoding game footage and facial expressions, with the facial expression network following a VGG16-like architecture [14]. Audio

[2]In this context, we consider that game stream highlights do not only depend on the game footage, but also on the social and emotional behaviour of the streamer.

features are extracted by using the principal components of Fourier coefficients, while subsequently a recurrent layer is used for fusion. The premise is that a high reconstruction error points to novelty, and novelty can be considered as a proxy for detecting interesting parts of the stream. This also deems the method suitable for *segmenting* a long stream into episodes that are likely to contain emotional and social content, especially when utilizing the streamer's face and voice. Evaluation was carried out on over 5 hours of footage broadcasted by two streamers playing a popular game. Results presented in [9] show that, as expected, analysis of facial expressions and audio significantly increases the precision of detected episodes, compared to simply using the game footage. In Fig. 3, we show some sequences from detected episodes where social and behavioural signals can be observed. In the first row, the sequence shows the player winning a game and reacting positively to this. In the second, the player is reacting to an in-game event. In the third row the player appears intrigued. Finally, in the fourth row the streamer is interacting with viewers.

Figure 3: Sequences from episodes extracted by applying [9] on face video, player audio, and game footage.

5 RESEARCH DIRECTIONS

Firstly, further studies should utilize a more rich and diverse set of streamers, while gathering enough game-specific data to facilitate modelling game context. Furthermore, several research directions span from exploiting such a vast source of heterogeneous data. For example, while unsupervised methods such as [9] are suitable for segmentation, supervised learning approaches can be employed to learn specific appraisals of the players emotional state directly from the multi-view observed data (e.g., by using latent dimensions such as valence-arousal [11]). Extraction of video-based physiological measurements [2] can also potentially contribute towards this. Learning mappings from a subset of views leads to further interesting problems, e.g. predicting player behaviour and experience directly by observing game footage raw pixels. Given appropriate data, complex behaviours such as *mimicry* and *conflict* can be studied in an audio-visual manner, by exploiting recent advances in machine and deep learning that scale to large datasets. Furthermore, semi-supervised methods can be used to model game-context as *side-information*, while creating platforms for recording other types of information (e.g., playtraces) can lead to further promising research directions.

6 CONCLUSIONS

In this paper, we discuss the underlying potential in utilizing the vast amounts of audio-visual data broadcasted during a typical game stream. We argue and show that the interactive setting under which a stream operates is enriched with emotional and social signals, thus providing a vast source of multi-view data that can be utilized towards tasks such as the machine analysis of social and emotional behaviour.

ACKNOWLEDGMENTS

We thank the NVIDIA Corporation for providing a Titan X GPU used in this work. Charles Ringer's work is supported by EPSRC grant EP/L015846/1 (IGGI). The authors would like to thank Mats Kathage (P4wnyhof) for permission to use footage from his streams.

REFERENCES

[1] Daniel Gábana Arellano, Laurissa Tokarchuk, and Hatice Gunes. 2016. Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games. In *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 184–193.

[2] Weixuan Chen and Daniel McDuff. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. *arXiv preprint arXiv:1805.07888* (2018).

[3] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis. 2015. The platformer experience dataset. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 712–718. https://doi.org/10.1109/ACII.2015.7344647

[4] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina. 2015. Audiovisual fusion: Challenges and new approaches. *Proc. IEEE* 103, 9 (2015), 1635–1653.

[5] Mehdi Kaytoue, Arlei Silva, Loïc Cerf, Wagner Meira Jr, and Chedy Raïssi. 2012. Watch me playing, I am a professional: a first study on video game live streaming. In *Proceedings of the 21st International Conference on World Wide Web*. ACM.

[6] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2018. Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *arXiv preprint arXiv:1804.10938* (2018).

[7] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[8] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM.

[9] Charles Ringer and Mihalis A. Nicolaou. 2018. Deep Unsupervised Multi-View Detection of Video Game Stream Highlights. In *International Conference on Foundations of Digital Games*. ACM.

[10] 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*. ACM, 36.

[11] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[12] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2015).

[13] Björn W. Schuller. 2018. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM* 61, 5 (April 2018), 90–99. https://doi.org/10.1145/3129340

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

[15] Georgios N. Yannakakis and Julian Togelius. 2011. Experience-Driven Procedural Content Generation. *IEEE Transactions on Affective Computing* 2, 3 (2011), 147–161. https://doi.org/10.1109/T-AFFC.2011.6

[16] Georgios N. Yannakakis and Julian Togelius. 2018. *Artificial Intelligence and Games*. Springer. http://gameaibook.org.