Streaming Behaviour: Livestreaming as a Paradigm for Analysis of Emotional and Social Signals

Charles Ringer*, Mihalis A. Nicolaou†

*Department of Computing, Goldsmiths, University of London, London
†Computation-based Science and Technology Research Center, The Cyprus Institute, Cyprus

***Abstract*—We argue that video game livestreams 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 human behaviour in groups, utilizing the 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.**

I. Introduction

Video game livestreaming 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` [1]. This paper is motivated by two relevant observations, namely (i) the inherently interactive nature of streaming [2], 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 mostly on 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.

II. Related Work

Most related to the livestream setting are databases collected towards player experience modelling [3]. For instance, the Platformer Experience Dataset (PED) [4] 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. The 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 [5], 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 [6], while wristbands are used for extracting physiological data - thus requiring access to the streamer. Works on analysis of affect and behaviour in co-operative and single player games include [7], where physiological and visual data from eight subjects is utilized, as well as [8], [9] where player experience modelling is utilized for tasks such as clustering player types. Finally, we refer the reader to recent articles such as [10], [11] 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 [12].

III. Streaming & Multi-View Data

In this section, we discuss data sources emerging in a livestream 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 nature of the stream points to research in machine learning (e.g., fusion [13]), computer vision, and acoustic

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

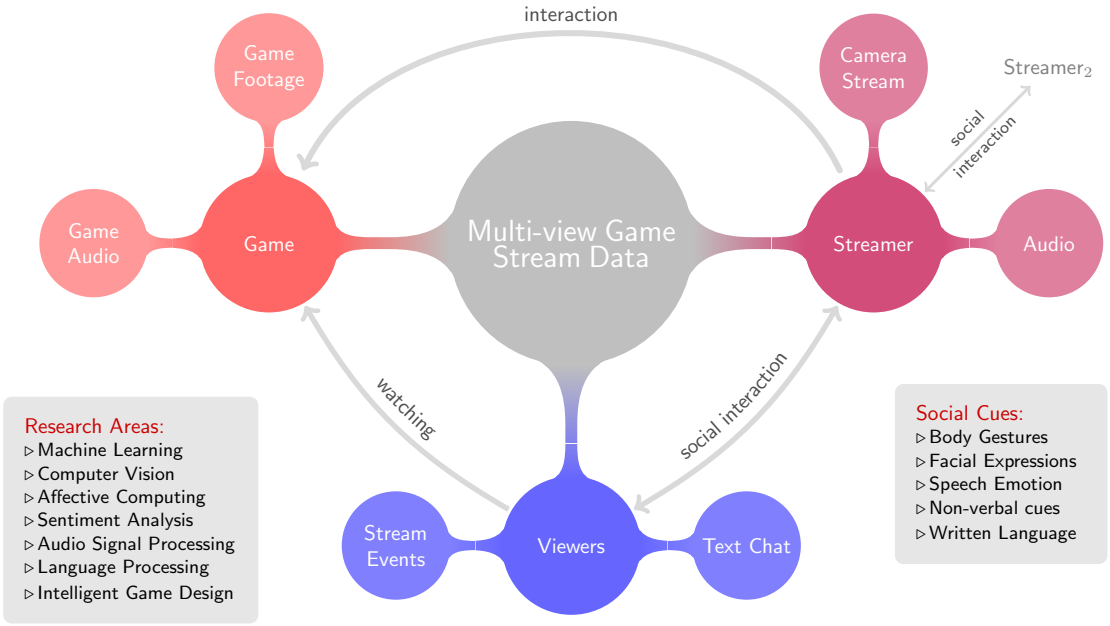

Fig. 1. Illustration of main concepts discussed in the paper, including social interactions, social and emotional signals conveyed, as well as related research areas. The streamer broadcast includes a video stream capturing face and/or body, along with audio recorded from the streamer's microphone. This 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 group behaviours via voice and facial expression analysis using data arising from both broadcasts, further informed by text-based communication with viewers.

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 challenging and rich data paradigm.

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

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 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 [14], [15].

IV. Results

In this section, we present two initial studies utilizing the aforementioned data sources. While stream data is in abundance, curating annotations for training machine learning models is a tedious, laborious, and expensive task that requires some form of human supervision, especially when annotating per-frame labels. It is therefore reasonable to firstly approach this problem in the context of unsupervised learning. In a recent work [16], we propose a deep unsupervised novelty-detection approach for detecting highlights in game streams[1], 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 [17]. Audio features are extracted by using

[1]Game stream highlights are considered to not only depend on the game footage, but also on the social and emotional behaviour of the streamer.

TABLE I
DISTRIBUTION OF CLASSES.

Valence			Arousal		Game Context								
Neg	Neut	Pos	Neut	Pos	In Lane	Shopping	Ret. to Lane	Roaming	Fighting	Pushing	Defending	Dead	Misc.
246	6,227	727	6,755	445	2,418	294	591	1,422	892	213	233	831	308

TABLE II
F1 SCORES FOR EACH LABEL ACROSS ALL MODELS. (V=VALENCE, A= AROUSAL)

Model	Task	Neg V	Neut V	Pos V	Neut A	Pos A	In Lane	Shopping	Returning	Roaming	Fighting	Pushing	Defending	Dead
Early Fusion	Joint	0.194	0.911	**0.362**	0.969	0.509	0.778	0.797	0.496	0.667	0.515	0.568	0.544	0.899
	Single	0.206	0.905	0.345	0.966	**0.540**	**0.842**	0.724	0.591	0.794	0.565	0.610	**0.667**	0.924
Late Fusion	Joint	**0.286**	**0.925**	0.297	0.964	0.465	**0.840**	**0.828**	**0.615**	**0.805**	**0.635**	**0.652**	0.557	0.906
	Single	0.088	0.918	0.340	0.968	0.491	0.791	0.776	0.513	0.774	0.581	0.582	0.452	0.937

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 show that, as expected, analysis of facial expressions and audio along with game footage significantly increases the precision of detected episodes (77%), compared to simply using the game footage (30%). 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.

Fig. 3. Sequences from episodes extracted by applying [16] on face video, player audio, and game footage.

Next, we present a study [18] on modelling both game context as well as emotional behaviour in a supervised manner. In this study, we collected a dataset[2] consisting of *League of Legends* using `TWITCH.TV`. The dataset includes 10 streamers (five male and five female), leading to a 10 hours of footage split in 7200 five-second long non-overlapping segments. Each clip was manually annotated both for game context, as well as for the emotional state of the player in terms of both valence and arousal [19], [20]. The network is similar to [16], following a stacked architecture of convolutional [21] and recurrent (LSTM [22]) layers, while further employing residual blocks [23]. In Table II, we show results on each of the emotion prediction tasks as well as game-context tasks. Both late and early-fusion are employed, while comparing on two tasks: (a) using a single model for emotion and game context, and (b) using a joint model that predicts both game context and streamer behaviour. As can be seen, results appear promising for several classes, especially for game context classification, with late fusion consistently outperforming early fusion. As expected, jointly modelling both game-context and player behaviour leads to better results. Nevertheless, accurately modelling the emotional state of the player remains a challenging problem, even when game context is considered, and these results can be considered baseline. This is made more difficult by the fact that some classes are naturally under-represented in game streams (see Table I).

V. 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 resource of multi-view data that can be utilized towards tasks such as the machine analysis of social and emotional behaviour. The interactive nature of game streams facilitates studying human behaviour in groups (streamers, viewers) and presents several challenges to researchers working in machine learning, computer vision, as well as signal and natural language processing.

[2]https://github.com/charlieringer/LoLEmoGameRecognition

REFERENCES

[1] M. Kaytoue, A. Silva, L. Cerf, W. Meira Jr, and C. Raïssi, "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, 2012.

[2] T. Smith, M. Obrist, and P. Wright, "Live-streaming changes the (video) game," *Proceedings of the 11th european conference on Interactive TV and video - EuroITV '13*, p. 131, 2013.

[3] G. N. Yannakakis and J. Togelius, "Experience-Driven Procedural Content Generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5740836/

[4] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis, "The platformer experience dataset," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept 2015, pp. 712–718.

[5] R. Robinson, Z. Rubin, E. M. Segura, and K. Isbister, "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, 2017, p. 36.

[6] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "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, 2016.

[7] D. G. Arellano, L. Tokarchuk, and H. Gunes, "Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games," in *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 2016, pp. 184–193.

[8] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[9] S. Asteriadis, K. Karpouzis, N. Shaker, and G. N. Yannakakis, "Towards detecting clusters of players using visual and gameplay behavioral cues," *Procedia Computer Science*, vol. 15, pp. 140–147, 2012.

[10] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, 2015.

[11] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018. [Online]. Available: http://doi.acm.org/10.1145/3129340

[12] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018, http://gameaibook.org.

[13] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[14] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[15] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, Jun 2019. [Online]. Available: https://doi.org/10.1007/s11263-019-01158-4

[16] C. Ringer and M. A. Nicolaou, "Deep unsupervised multi-view detection of video game stream highlights," in *International Conference on Foundations of Digital Games*. ACM, 2018.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] C. Ringer, J. A. Walker, and M. A. Nicolaou, "Deep unsupervised multi-view detection of video game stream highlights," in *2019 IEEE Conference on Games*. IEEE, 2019.

[19] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[20] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120 – 136, 2013, affect Analysis In Continuous Input. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885612001084

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," vol. 9, pp. 1735–80, 12 1997.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.