Finding the Interesting Moments in Video Game Livestreams Through Deep Learning

Charlie Ringer | University of York | cr1116@york.ac.uk | @charlieringer | charlieringer.com

WHY SHOULD YOU CARE?

Livestreaming has exploded in popularity recently. Twitch.tv, generates over 132 *years* of content every day. It offers us powerful insights into how a subset of players (streamers) are interacting with games. Finding the most 'interesting' moments in these streams is attractive to:

- **Streamers:** Highlights are an alternative content stream, shared through services like YouTube.
- **Viewers:** Viewers are able to consume the 'best bits' of more of their favourite streamers.
- **Game Developers:** Interesting moments in a game can inform future game design.
- **Streaming Platforms:** Interesting moments are advertising, and can entice potential viewers.
- **Researchers:** Organic environments are key to advancing our ability to model the real world.

Fig 1. 1st row: various facial expressions. 2nd row: various body gestures. 3rd row: various head poses. 4th row: challenging frames with facial occlusions. (thanks to p4wnyhof for permission to use his stream data).

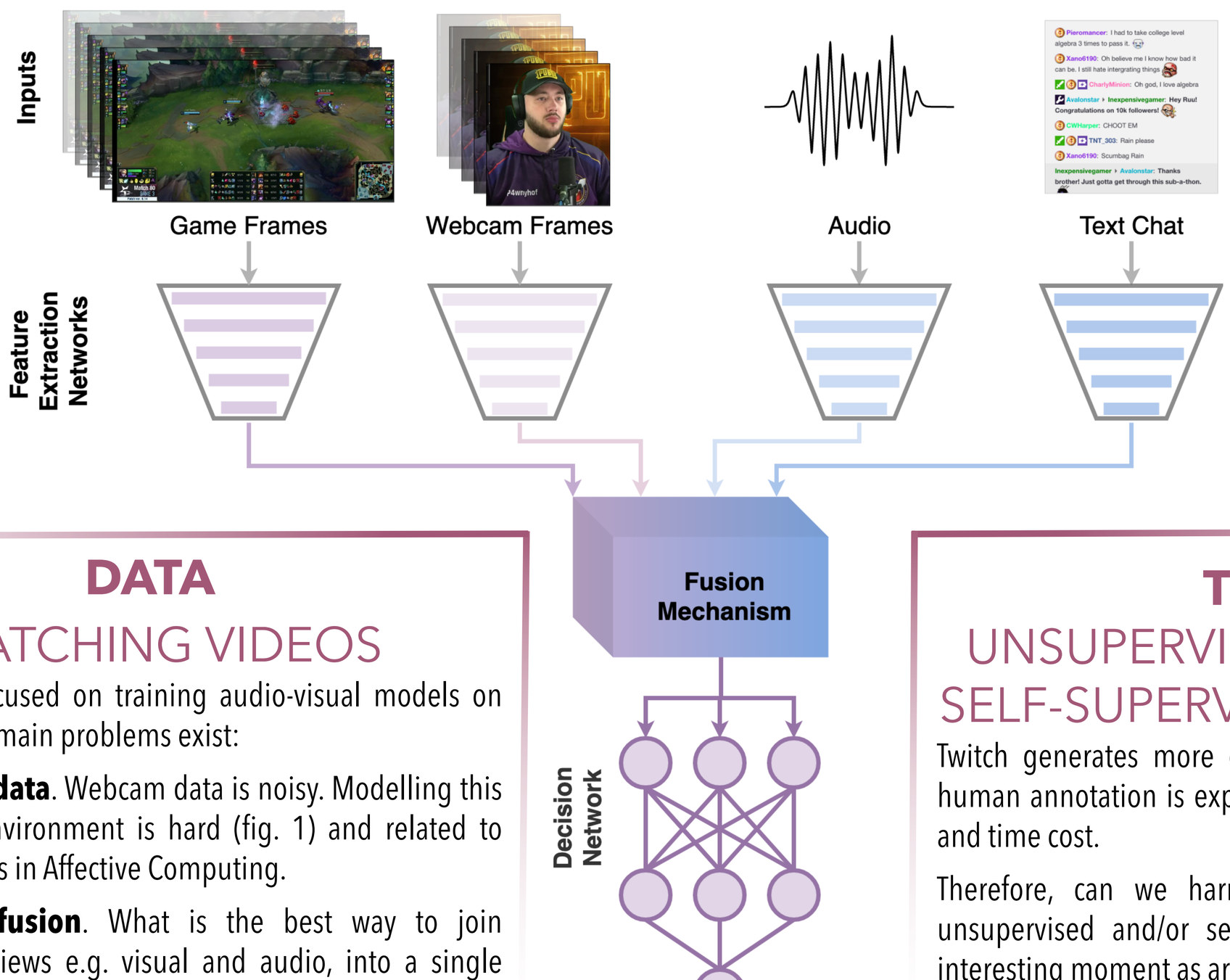

Fig 2. An overview of a potential system proposed in this research.

1. First features are extracted from each raw data view (game, face, audio, text).
2. Next, these features are joined in a fusion process.
3. The fused features are then passed into a decision network which classifies the inputs as either part of a interesting moment or not.

DATA

WATCHING VIDEOS

Initial works focused on training audio-visual models on video data. Two main problems exist:

'In-the-Wild' data. Webcam data is noisy. Modelling this uncontrolled environment is hard (fig. 1) and related to many challenges in Affective Computing.

Multi-modal fusion. What is the best way to join multiple data views e.g. visual and audio, into a single unified model? (fig. 2)

READING CHAT

During livestreams the audience can interact with the streamer in real-time through written chat, providing an extra data view, unique to livestreams.

The main goal here is to transform a word (or a set of words) into a vector (a series of numbers) in a sensible way where similar words have similar vectors.

These vectors can then be used in the same way as image and audio data in a deep neural network (fig. 2).

TRAINING

UNSUPERVISED LEARNING? SELF-SUPERVISED LEARNING?

Twitch generates more data than can be processed but human annotation is expensive, both in terms of financial and time cost.

Therefore, can we harness this huge data alongside unsupervised and/or self-supervised learning to model interesting moment as artifacts of the data itself?

GANS?

Generative Adversarial Networks (GANs) can create new media, usually from a random input. What if, instead of a random input, we use long, unedited, livestream videos. Can we train a GAN to pick out the best parts?

PUBLICATIONS

Interested in this work? My publications can be found at charlieringer.com/publications or by scanning the QR Code.

Supervisors: Dr. James Alfred Walker & Dr. Mihalis A. Nicolaou

EPSRC Engineering and Physical Sciences Research Council

iGGi Your future in games research

University of York