Summarizing E-Sports Matches and Tournaments

The Example of Counter-Strike: Global Offensive

Mathias Lux
Alpen-Adria-Universität Klagenfurt, Austria

Pål Halvorsen*
SimulaMet, Norway

Duc-Tien Dang-Nguyen
University of Bergen, Norway

Håkon Stensland†
Simula Research Laboratory, Norway

Manoj Kesavulu
Dublin City University, Ireland

Martin Potthast
Universität Leipzig, Germany

Michael Riegler
SimulaMet, Norway

ABSTRACT

That video and computer games have reached the masses is a well known fact. Furthermore, game streaming and watching other people play video games is another phenomenon that has outgrown its small beginning by far, and game streams, be it live or recorded, are today viewed by millions. E-sports is the result of organized leagues and tournaments in which players can compete in controlled environments and viewers can experience the matches, discuss and criticize, just like in physical sports. However, as traditional sports, e-sports matches may be long and contain less interesting parts, introducing the challenge of producing well directed summaries and highlights. In this paper, we describe our efforts to approach the game streaming and e-sports phenomena from a multimedia research point of view. We focus on the challenge of summarizing matches from specific relevant game, Counter-Strike: Global Offensive (CS:GO). We survey related work, describe the rules and structure of the game and identify the main challenges for summarizing e-sports matches. With this contribution, we aim to foster multimedia research in the area of e-sports and game streaming.

CCS CONCEPTS

• **Human-centered computing** → *Virtual reality*;

KEYWORDS

E-sports, video summaries, multimodal, Counter-Strike

ACM Reference format:
Mathias Lux, Pål Halvorsen, Duc-Tien Dang-Nguyen, Håkon Stensland, Manoj Kesavulu, Martin Potthast, and Michael Riegler. 2019. Summarizing E-Sports Matches and Tournaments. In *Proceedings of 11th International Workshop on Immersive Mixed and Virtual Environment Systems, Amherst, MA, USA, June 21, 2019 (MMVE'19)*, 6 pages.
https://doi.org/10.1145/3304113.3326116

*Also with Oslo Metropolitan University, Norway.
†Also with University of Oslo, Norway.

MMVE'19, June 21, 2019, Amherst, MA, USA

ACM ISBN 978-1-4503-6299-3/19/06…$15.00
https://doi.org/10.1145/3304113.3326116

1 INTRODUCTION

E-sports is huge. Already in 2013, the number of concurrent users for a single event exceeded eight million for a League of Legends Championship[1]. In 2016, approximately 162 million viewers accessed e-sports streams frequently[2]. The rich bouquet of data including audio and video streams, commentaries, game data and statistics, interaction traces, viewer-to-viewer communication, and many more channels allow for particularly challenging multimedia research questions. From an observational point of view, it has been argued that the interactivity of participating in a game affects the narrative of the game through the player's choices. While this may be the case for single player games, the case of e-sports matches is much more complex. The game provides a sandbox with a very limited toolkit and two groups of people competing over a virtual price. Arguably, this might be a new version of the medieval jousting or the Roman gladiators, but the new aspect is that everyone, everywhere can offer or enter an arena and share a view of the game to everyone else. This has become so popular that it is even integrated in the current hardware generation of game consoles. Gamers can stream to Twitch, YouTube, or Mixer at the press of a button from XBox One and PlayStation 4. As a result, an immense amount of data is streamed and offered to and from users. Many of the viewers face the problem of trying to sift through this vast amount of data offered to find those bits and pieces most relevant to them. A viewers interface can be found in Figure 1. This is already an advanced view, where viewers have to configure their view and use the timeline to navigate through the videos. While this being a traditional information retrieval problem, the technical and cultural context is new. Interactive entertainment has evolved beyond controlled channels and broadcasters to user generated content, which is unfiltered and raw. Furthermore, typical characteristics include that

(1) the interactivity is extremely high,
(2) a lot of the footage is multi-view content, where many concurrent players interact in the same game world instance, and

[1] https://associate.vc/esports-millions-of-viewers-millions-of-dollars-e7b411b57ba6
[2] https://www.statista.com/statistics/490480/global-esports-audience-size-viewer-type

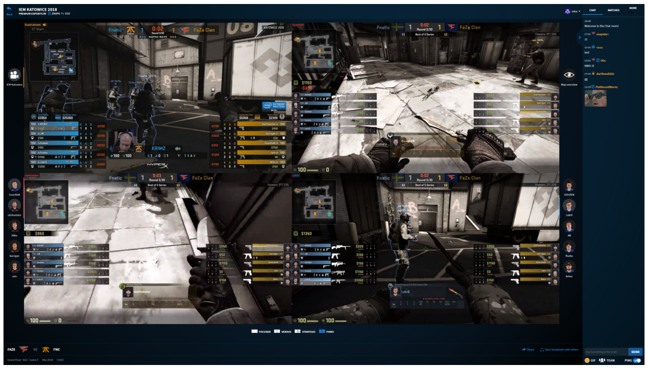

Figure 1: Screenshot of an interactive view of an ESL One CS:GO match. Four video streams can be displayed, and viewers can configure the layout and which players to show. A chat window is on the right.

(3) the user group is typically well-informed, very similar to core sports fans who know every piece of history from their soccer, hockey, or American football team.

The increased interactivity through chats between viewers as well as from audience to players encourages intense discussions, but also trolling, the use of bots, hate speech, toxic behavior, etc.

In this paper, we address one of the key challenges of the industry today, which is to render the results of e-sports events in a way that makes them more promotable to the audience. It is notoriously difficult to search, e.g., for exciting highlights, because of the huge amounts of video recorded for each match, and their homogeneity, e.g., every match of League of Legends roughly looks the same. A lot depends on the manual selection and curation of supplementary material for promotion on an event's website. This way, individual matches of e-sports events can be more easily accessed by those attending on-site or by stream, however, at the cost of significant manual overhead. At the same time, these efforts also enable the event and its highlights to be retrieved or recommended later on. In this work, we use *Counter-Strike: Global Offensive* (CS:GO), having a solid player and viewer base, as a case study. CS:GO serves more than 10,000 matches simultaneously with over 240,000 players[3]. We investigate the game's structure as well as the typical e-sports match setup and how much data is generated throughout a match or a series of matches in a tournament. We describe the task of summarizing matches and tournaments, and we identify and list the challenges for automatic summarization to bring this problem to the awareness of researchers. In addition, we present results from the MediaEval 2018 Gamestory task which was focused on summarizing games in a best possible way. The subjective evaluation shows that there are potentials, but still room for improvements.

2 BACKGROUND AND RELATED WORK

E-sports is multimodal. It contains a wide range of modalities including *video streams* from multiple perspectives on a given match, *audio streams* from players, commentators, and moderators, *telemetry* obtained from game engines, and real-time *viewer feedback* in implicit and explicit forms via user interactions, chat, and social media. Even streams of health data of players, such as their heart rates, may be available. At the same time, a summary of an e-sports event will be multimodal as well, involving the traditional kinds of media used for news coverage and promotion, namely text, imagery, and video. Naturally, the task of automating the summarization of an e-sports match will have to draw the expertise of the outlined fields. This section collects examples of closely related approaches from three selected fields, namely sports, natural language generation, and video summarization. In traditional sports, there have been numerous approaches to detect events of interest within matches that may form the basis of a summary: for instance, in soccer, by combining audio, shot boundary and classification, and game restarts at the center of the field [3], and in baseball, by combining video and audio [18]. A survey of multiple sports-related event detection approaches is presented in [13]. In e-sports, the most closely related work is found within game analytics and game data mining, where the goals typically involve user research, analyzing user experience, and the definition of related metrics. An exhaustive overview of the field has been compiled by El-Nasr et al. [2], actively run competitions on game analytics data are for instance hosted on Kaggle[4]. Based on this body of work, the quantitative data required to characterize an e-sports match will be derived directly from within a game engine. In this context, game engine developers already integrate sophisticated analytics technology for the purposes of improving the games. The field of natural language generation has tackled the generation of news articles from quantitative data for decades, including the generation of sports news (e.g., [12]). Owing to the profound difficulties associated with generating natural language that prevail to this day, approaches are mostly rule-based or template-based. Application domains besides sports include weather forecasts, financial reporting, and other related domains where the text written mostly depends on data. Few such systems have come to fruition as of yet [11]; nevertheless, recent startups such as Narrative Science or Automated Insights have gained some traction in selling their automated news coverage of sports and other fields. We believe that e-sports summarization will be a natural extension of these efforts, since much like traditional sports, rich and fine-grained data is available via game telemetry and game analytics. Various studies are already assessing whether automatically generated news are accepted by readers (e.g., [4]), finding that human subjects do not rate human-written news better in all respects than computer-written news, or vice versa. Dörr [1] surveys the market of algorithmic journalism in more detail.

Regarding video (and image) summarization, the multimodal story-oriented video summarization (MMSS) [9] approach integrated multimodal information (textual terms, scene frames, and logo information) in a graph, treating it in a uniform, modality-independent fashion, with no need of parameter tuning. In the area of multimodal or cross-modal retrieval using deep neural networks, Lei et. al. [5] proposed a temporal segmentation framework based on clustering of both visual and semantic affinity graph of the video frames. A pre-trained deep convolutional neural network (CNN) extracts deep visual features constructing a visual affinity graph, and a semantic affinity graph is constructed based on word embedding

[3]https://csgo-stats.com/, accessed 2019-02-12

[4]https://www.kaggle.com/tags/video-games, last accessed 2019-02-20

of the frames' semantic tags generated from an automatic image tagging algorithm. A dense neighbor method is then used to cluster the joint visual and semantic affinity graph to divide the video into sub-shot level segments and from which a summary of the video can be generated. Moreover, the deep side semantic embedding (DSSE) model [19] finds semantic meaningful frames or shots of videos with the help of side semantic information. DSSE makes a latent subspace by correlating hidden layers of two uni-modal autoencoders (video frames and side information). The semantic relevance is effectively measured, and video segments by minimizing the distances to the side information in the constructed latent subspace. The cross-media multiple deep network (CMDN) model [10] was used to retrieve images based on the images and corresponding text where the learning stage first used multimodal Deep Belief Network to model the inter-media separate representation and Stacked Autoencoders to model the intra-media separate representation for each media type. Then, a two-level network, including the joint Restricted Boltzmann Machine and Bimodal Autoencoders, was used to get the final shared representation for each media type of the cross-media data. Furthermore, Ngiam et. al. [8] discussed multimodal deep learning in order to allow the network to learn features from multiple modalities for audio-visual speech classification. Moreover, for learning a joint representation of multimodal data, multimodal deep belief networks [14] define a probability distribution over the space of multimodal inputs and allow sampling from the conditional distributions over each data modality, i.e., creating a multi-model representation.

It is self-evident that for the task of e-sports summarization, we may build on the aforementioned approaches exemplified. However, many of these approaches only partly match our scenario, e.g., retrieving one modality using another. Rather than targeting a single modality as a result, a truly useful game summary will incorporate text, imagery, and video as well as appropriate statistics, composed into a coherent news-like article or video, offering new challenges for summarization research.

3 CS:GO - THE GAME

Narratives in video games are typically modeled by game designers, but the most engaging games are those, where the interaction of gamers actually creates the story. This is especially the case with games played for e-sports. With CS:GO for example, the context is clearly defined: in one of the game modes (DE_Map), terrorists try to place a bomb and counter terrorists try not to let them. While the same general story plays out over and over again in diverse matches, the viewers appreciate the evolution of a match, the tactics applied, how those affect the outcome, the ramping up to a critical point and the tension at the tipping point of the match. Like in popular sports, i.e., soccer, this is very similar to a unique story build from the same narrative elements used all over again.

CS:GO is a very good starting point as there are solid player and viewers bases, well organized tournaments, professional teams and leagues. Game mechanics and players involved are easily understandable and manageable. This, however, is just a beginning as ever expanding games with additional degrees of freedom enter the scene in fast succession. From CS:GO with a very limited set of possible actions and a small number of players, one could extend to the battle royale genre, where scarce resources, first person shooter mechanics, and tower defense are combined on servers with 100 concurrent players per map instance. Tension is built artificially by making the map smaller and smaller to lead the players to a final stand off.

CS:GO as an e-sports game has very specific rules. First of all, matches consist of several rounds, and players do only re-spawn in between rounds. So, if a player's avatar gets eliminated, the player has to wait until the next round begins. In between rounds, players can outfit their avatars with weapons, ammo, tools, and armor. The resources available for outfitting are determined by the success in previous round, e.g., players earn money if they plant or defuse the bomb, eliminate other avatars, and survive a round. The economy of money available follows a complex rule set[5] also involving the success history (or lack thereof).

A typical CS:GO match is decided in a best of 30 fashion, i.e., the first team scoring 16 rounds wins the match. After 15 rounds team switch sides, i.e., terrorists (T) become counter terrorists (CT) and vice versa. If two teams end up with a draw after 30 rounds, rules specific to the tournament or league (context of the match) apply. Typically, draws are accepted in leagues and stages in which teams play for points, but not in single elimination stages, where teams then play overtime (typically in 3-5 rounds). Outfitting the avatars and planning for in-game money (economy management) are critical parts of a team's strategy in competitive CS:GO matches. Strategies typically span over multiple rounds and include *full buys*, where teams can afford to buy what they need to win the round, *eco rounds*, where teams save money for full buys, *anti-eco rounds* as a counter strategy to eco rounds, and so on. In general, the goal of each round is to eliminate all opponents. However, there are also map-based goals to win a round. For example, beside to eliminate all opponents, in a DE_Map round, the goal for the T team is to plant the bomb to some specific locations and protect it from being defused (until the end of the round), while CT team has to defuse the bomb (if it is already planted) before detonation (typically in 40 seconds). Other map types and summaries of the CS:GO game plays are illustrated in Figure 2.

Tactics within rounds are different for the two sides. For example, in DE_Map game mode, teams playing the T side in the beginning of a round often spread out and try to prevent early aggression from the CT side. They wait for smoke and other grenades the CT side might set if they had enough money. In the middle of a round, they often flank the CT to gain intelligence on the positions held by the CT side even moving as far as the CT spawn locations to sneak up on the enemy from behind. They set up for going active, i.e., rushing into the critical area and planting the bomb, and eventually position one team member as a dedicated sniper for cover. At the end of a round, the players execute their plan by planting the bomb and defending positions while the players flanking try to cut off the rotations (position changes) on the CT side.

Teams playing the CT usually try to prevent players from the T side from reaching areas where they can plant the bomb, or at least delay them. The CT side typically starts out by throwing a batch smoke or flame grenades. First objective is to counter early rushes, but also to gain control over strategic points on the map

[5] http://www.tobyscs.com/csgo-economy-guide/, accessed 2019-02-12

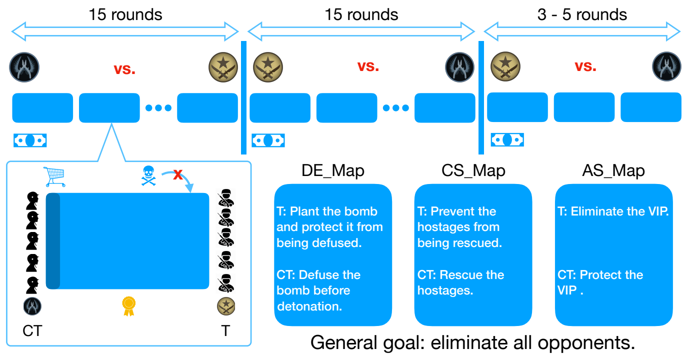

Figure 2: Overview of general CS:GO game play. Two teams, terrorists (T) and counter terrorists (CT), play *DE_Map* (majority), *CS_Map*, or *AS_Map* (less common than the others), being the game modes played in e-sports in a best of 30 fashion.

played by early aggression, i.e., forcing players from the T side into shoot outs and stand offs. In the middle of the round, players try to find out where the T side tries to plant the bomb and to delay the action of the T side as much as possible with grenades. Depending on the map there are multiple points where this is possible. At the end of the round players try to stop the execution of the plan from the T side or, if not possible, fall back, regroup and try to go in and defuse the bomb.

The strategy becomes even more interesting in CS_Map game mode, in which CT teams have to rescue any of the hostages (typically two hostages for each round in CS:GO version) and lead the hostages to the Hostage Rescue Zone (HRZ) / Escape Zone (EZ). Choosing the right strategy for a team in this mode is more challenging than in the DE_Map mode since hostages are spawned generally near the area, where the T side starts the round. There is no room for strategies such as "faking" bomb-sites since you can target either (or both!) hostages. Moreover, time plays a different role, as it depends on players actions, i.e., the game clock is extended by one minute after hostage pick up even if the player doing the pick up is immediately eliminated.

Professional players and teams in CS:GO heavily rely on voice communication, trust and adaptability to new situations. Team members have to inform the other on positions of enemy players, their current status and their individual plans and moves. Team members have to trust each other that everyone is doing what's best for the team, e.g., that players with an abundance of money will drop weapons for other teammates or players from the T side will hand over the bomb to a designated player.

As a match in CS:GO develops over multiple rounds the situations are always different. Teams have to adapt to new situations in short time to be able to turn the game around to their favor. Possible tipping points in CS:GO tournaments are rounds where both teams can afford to buy what they need. These are time points when teams can go for comebacks after consecutively loosing rounds to the other team. Thus, as can be understood from the information above, the gameplay is complex, and making summaries of CS:GO matches raises several challenges.

4 CHALLENGES FOR SUMMARIZATION

Compared to sports summaries, computer and video games are not focused on a small number of attention points, like, for instance the ball and the goals on a soccer field, but provide multiple views and concurrent events within a clearly defined game map, e.g., 10 players and a commentator's view in CS:GO. Compared to a soccer match, there are also clearly defined and well-formatted statistics on events relevant to the game's progress within these streams. So, it is not only a linear outplay, but often many things happen (nearly) at once, and all together, they lead to a specific outcome. Also, events may be connected across longer time frames, for example, one player setting a booby trap at the beginning of a match, whereas another player runs into it only at end, deciding the game. Here, the importance of setting the trap and, therefore, taking note of it is hence determined by whether it has an influence of the outcome of the game (or whether it is "funny" to see someone stumbling into it). Moreover, game statistics only carry active and obvious events, but they miss those events with semantics on a tactical level, like intentional misses, lures, fake tactics, intentional acts risking a respawn or player's avatar death, etc.

In CS:GO, a good summarization should be able to reflect the development of a match over all rounds by showing the economy management and the effect of it on the rounds played. Tipping points of a match economy wise should result in turning around consecutive fails in consecutive wins. However, in contrast to the economy management, tactics employed within the round also impact the outcome in combination with the execution of the economy management. Lucky shots, badly synchronized plan execution, or even bad luck can change the game. There are several challenges we already identified for summarizing CS:GO matches:

Player positions Tactics within the rounds heavily rely on where players are within the map. For CS:GO and many other games, the position of players in the map is not given in the meta data stream, but can only be inferred visually from an overview map. Moreover, the map is not flat, but often has multiple floors and points of interest may lie on the same x and y coordinates, but at a different z.

Classification of intra-round strategy Depending on economy management, teams employ different strategies for a single round, like full buy, eco, or a general strategy of disruption. For summarization, it is useful to know which team employed which strategy for a given round.

Classification of inter-round strategy Teams also have an overall strategy, like *wait and see* for actions of the other team, or *rush in* to try to win the match as soon as possible.

Tipping points and events Beside the knowledge of who won the match, viewers also appreciate to reflect on the most critical time points in the game, like for instance the goals, fouls, or corner throws in a soccer game. This includes events when players fail to implement a strategy, when teams turn around the game after consecutive looses, etc.

Ranking of player importance A team can only be as good as its weakest player. For summarization, we need to identify the relevance of contribution to winning or loosing for each player. This goes far beyond elimination count and includes team play, communication, skill and much more.

Evaluation of summaries While it is easy to create a short video from multiple streams and present it to viewers, the viewers ultimately decide if the summary was good or not. There are a few hard requirements for summaries, but a lot of soft ones. One has to think of an objective way to judge the precision and recall of summaries.

Although the challenges discussed here are for game that is not commonly played in virtual reality (VR) there is a trend in e-sports to go towards VR tournaments[6]. This is more focused on the games played at the moment, but also streaming of games as VR streams can be a future option. In that case, summarization in its basics stays the same, but complexity if added in terms of VR. Questions that arise there are for example (i) what is a good view point, (ii) should the viewer control the view or follow the players movement, (iii) will viewers get motion sickness from following the fast movements of professional players, and many more.

5 GAMESTORY 2018 TASK RESULTS

At MediaEval 2018, a novel task focusing on e-sports summarization was introduced. The goal for each submitting team was to provide a summary of a CS:GO game [6]. As described above, a good summary should be able to reflect the development of a match over all rounds by also showing the economy management and the effect of it on the rounds played. Three to four judges from a jury reviewed submissions and ranked them.

5.1 Game Data

For the GameStory task, our partners from ZNIPE.tv provided training and test data from an ESL One tournament in Katowice in 2018. ESL is to date the largest e-sports organizer world wide. The ESL One tournament series focuses on premier offline tournaments of specific games, including CS:GO and DOTA 2, and counts among the most prestigious events in e-sports for CS:GO, also supported by the publishers and developers of CS:GO, the Valve Corporation. The data consists of twelve saved video streams along with meta data. Ten files give the view of the players with the in-game audio streams. One file gives the commentator stream, where a professional cutter selected the parts of the player views to be shown as well as videos from the audience and the teams and the commentator provides the spoken content mixed with recordings from the game and audience cheering. The last one shows the game world from above with icons indicating the position of players. A meta data file indicates the start and the end of the games and content in the commentators stream. JSON files (one for each match in the data set) capture player activity and events in addition to the raw video. Events range from kills, death, round end and starts to what the players bought at the beginning of rounds and when a if a bomb was set or a grenade was thrown or went off.

[6]https://vr.eslgaming.com

5.2 Summarization Approaches

For the task, which was run the first time in 2018, three teams submitted their approaches. The first approach [7] (approach 1) was based on the metadata, and there only on the economy actions, the round end and the kill streaks. This approach was rather limited and more interesting summaries could have been achieved by including for example other information like how the game ended. Another weak point was that videos were cut based on time stamps only which lead to cut off audio which was received negative.

The second approach [17] (approach 2) was based on automated killstreak extraction. The killstreak were identified using metadata and content based methods. The main finding was that the timestamp of the metadata were not useful and the authors relied on content based methods only. The results of the content based method were merged into a summary video with a multi-view perspective of the event stream, the actor's view and the views of the victims. The accuracy was almost frame accurate which made the summaries very smooth.

The final approach [16] (approach 3) was based on events. For the presented method, the gameplay was divided into several sequences. These sequences then were ranked based on to how many interesting event they were connected. Each event belonged to one of ten categories (spanning from assist to throw). The authors focused on kill events and match progress using the sequences to obtain the final weight which was then used to create a ranked time line of the match summary.

5.3 Evaluation

As there is not really a true correct summary, and the quality of the summary is based on people's opinions, we made a subjective evaluation using a jury. The jury, made up from experts from ZNIPE.tv, CS:GO players and game researchers, were asked to evaluate and reflect on the summaries. The jury outlined strong and weak points of the submissions and ranked them according to the summaries' ability to reflect the story of the match or tournament. The judges were asked to summarize both strong and weak points of submissions and to rate it using a 5-point Likert scale (strongly agree to strongly disagree) on the following statements:

(1) The submission gives a summary of the match at hand,
(2) The submission is entertaining,
(3) The submission provides the flow and peak of a good story,
(4) The submission provided an innovative way to present a summary of an CS:GO match and
(5) A summary like this submission can be applied to games different from CS:GO.

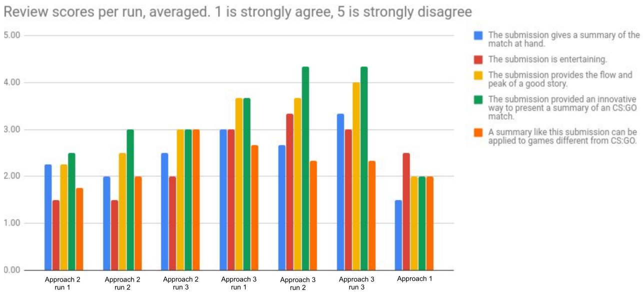

Figure 3: Jury evaluation of the three approaches. The scores are averages of the jury members individual scores.

Figure 3 shows a plot of the jury's evaluation of the various approaches and their respective submissions (runs). Approach 1's video summary var rated slightly better than the two others, and was according to the jury presented in an innovative way with a mix of different views. The second approach got the best entertainment scores, and the jury thinks that this is the best approach to be used for other types of games different to CS:GO. However, making an e-sports match summary is a complex task, and there are a large number of considerations to be made. According to the jury's evaluation, there are still room for much improvement. Neither of the suggested approaches gives a "perfect" summary of the game, but combinations of them might provide even better results. Nevertheless, more research is needed to find better approaches for summarizing e-sports matches. It is hard to make clipping accurate, select the appropriate events and at the same time take into account game strategies and various game events and their views.

6 CONCLUSION

Many people believe that interactive entertainment in the form of video games is here to stay. In a sense, one can argue that video games delivered what interactive TV promised all along. Video games provide a way to not only experience what the game designer and developers intended the audience to experience, but also allow players to shape the experience to make it individual [15]. However, recent developments have shown that there is a broad audience focusing on watching other people play through consuming game streams. Just looking at the popular web page Twitch.tv the average number of viewers has risen over 2018 and has exceeded 2.4 million viewers at peak view times in February 2019[7]. So, in our opinion, game streaming and e-sports have already found a way into everyday life of many people and is worth investigating.

With our work, we lay grounds for content analysis and multimodal summarization in the field of e-sports and streaming of e-sports. The example of CS:GO outlines the complex structure and the necessary understanding of the domain, which differs from game to game. However, the way to approach it stays the same. First, analyze the game rules, then analyze the emergent behavior of the players, then investigate the data at hand and, finally, develop an evaluation routine for the use case at hand.

Along the way, a lot of challenges have been identified. In our case, streams had to be synchronized down to frame level to be able to play them side by side. Although time stamps for the events shown in the streams were given in the meta data, the actual events were off up to 40 seconds plus or minus. As for the evaluation, the judges noted that submissions did not focus enough on the overtime of the CS:GO match at hand. They missed out highlights on the emotional peaks at these final rounds. However, the current work puts focus on an interesting research topic, and in the future, enhancements and combination of the proposed approaches should be tested.

[7]https://sullygnome.com/viewers, last accessed 2019-02-22

REFERENCES

[1] Konstantin Nicholas Dörr. 2016. Mapping the field of Algorithmic Journalism. *Digital Journalism* 4, 6 (2016), 700–722. https://doi.org/10.1080/21670811.2015.1096748

[2] Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. 2013. *Game Analytics: Maximizing the Value of Player Data*. Springer Publishing Company, Incorporated.

[3] Farhad Bayat Farshad Bayat, Mohammad Shahram Moin. 2014. Goal Detection in Soccer Video: Role-Based Events Detection Approach. *International Journal of Electrical and Computer Engineering (IJECE)* 4, 6 (Dec 2014), 979–988.

[4] Andreas Graefe, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19, 5 (2018). https://doi.org/10.1177/1464884916641269

[5] Zhuo Lei, Ke Sun, Qian Zhang, and Guoping Qiu. 2016. User Video Summarization Based on Joint Visual and Semantic Affinity Graph. In *Proc. of ACM iV&L-MM*. 45–52. https://doi.org/10.1145/2983563.2983568

[6] Mathias Lux, Michael Riegler, Duc-Tien Dang-Nguyen, Marcus Larson, Martin Potthast, and Pål Halvorsen. 2018. GameStory Task at MediaEval 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[7] Mathias Lux, Michael Riegler, Duc-Tien Dang-Nguyen, Marcus Larson, Martin Potthast, and Pål Halvorsen. 2018. Team ORG@ GameStory Task 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[8] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proc. of ICML*. 689–696.

[9] Jia-Yu Pan, Hyungjeong Yang, and C. Faloutsos. 2004. MMSS: multi-modal story-oriented video summarization. In *Proc. of IEEE ICDM*. 491–494. https://doi.org/10.1109/ICDM.2004.10033

[10] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In *Proc. of IJCAI*.

[11] Ehud Reiter. 2010. *Natural Language Generation*. Wiley-Blackwell, 574–598. https://doi.org/10.1002/9781444324044.ch20

[12] Jacques Robin and Kathleen McKeown. 1996. Empirically Designing and Evaluating a New Revision-Based Model for Summary Generation. *Artif. Intell.* 85, 1-2 (1996), 135–179. https://doi.org/10.1016/0004-3702(95)00125-5

[13] H. C. Shih. 2017. A Survey on Content-aware Video Analysis for Sports. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 5 (2017), 1212–1231. https://doi.org/10.1109/TCSVT.2017.2655624

[14] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Learning representations for multimodal data with deep belief nets. In *Proc. of ICML Representation Learning Workshop*.

[15] Tynan Sylvester. 2013. *Designing games: A guide to engineering experiences.* " O'Reilly Media, Inc.".

[16] Ninh Van-Tu, Le Tu-Khiem, and Tran Minh-Triet. 2018. GameStory: An event-based Approach. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[17] Michael Wutti. 2018. Automated Killstreak Extraction in CS: GO Tournaments. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[18] Min Xu, Ling-Yu Duan, Changsheng Xu, M. Kankanhalli, and Qi Tian. 2003. Event detection in basketball video using multiple modalities. In *Proc. of the joint International Conference on Information, Communications and Signal Processing and Pacific Rim Conference on Multimedia*, Vol. 3. 1526–1530 vol.3. https://doi.org/10.1109/ICICS.2003.1292722

[19] Y. Yuan, T. Mei, P. Cui, and W. Zhu. 2017. Video Summarization by Learning Deep Side Semantic Embedding. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 1 (2017), 226–237. https://doi.org/10.1109/TCSVT.2017.2771247