Cut Detection Survey and Analysis for Long Form Video Game Streams
Prof. Prosenjit Bose

Jaime Herzog

101009321

jaimeherzog@cmail.carleton.ca

April 5, 2020

1   Abstract

Shot Boundary Detection, or simply Shot Detection, is a fundamental part of research in the broader field of video analytics, used for essential video analysis tasks such as video indexing and content-based video retrieval. For professional video game live streamers, who create hours of continuous content with significant downtime, identifying cuts in their streams is an important first step for automatically generating condensed stream highlights. In this project, I have summarized the nomenclature and methodologies established in the academic canon for Shot Detection research, implemented a sample of the most common approaches, Colour Histogram comparison and Edge Change Ratio using Canny Edge Detection, and compared their effectiveness when used on video game live stream content, as well as establishing methodological challenges to this new content form.

2   Acknowledgements

Contents

3   Motivation

3.1   Why Shot Boundary Detection?

With the proliferation of internet connectivity in recent years, there has been a massive increase in the volume of video on the internet. With this growth, an industry built around the creation, hosting and sharing of video content grew as well, with YouTube alone having an annual revenue of over US$15 billion dollars [5]. Portable devices increase the accessibility for the creation and uploading of video content, and the combination of the massive social network surrounding content creation, with the enticing entertainment industry creating multi-millionaire content creators has exploded the sheer hours of video being uploaded daily. For YouTube, this type of content tends towards short-form video content at on average 10 minutes long. Comparing this to video game live streamers, who in some cases are live 8 hours a day, 5 days a week, and in some cases even more frequently, the volume of video uploaded per creator is significantly increased. In Q4 2019, there were 2.3 billion hours watched on Twitch, and 0.9 billion on YouTube Gaming [3].

For live streamers, when they have so much continuous uptime, it is important to create highlights of their daily streams, in order to increase their reach to viewers who don't have the time to watch their full time streams, as well as to diversify their brand onto a traditional YouTube channel. The creation of these highlights is extremely time consuming, as the hours of content must be sifted through manually by either the content creators themselves, adding hours to their daily workload, or by hired editors. Commonly, it is unfeasable for streamers to spend all their spare time manually editing highlight videos for their daily streams, and for streamers with a lower viewerbase and thus less income, it is also difficult to afford a full time editor. This creates another barrier of entry to the career of live streaming, a career track that is already difficult to get started in due to the high viewership required for career viability. The solution for this currently manual labour is an automated editing system, which methodologically identifies key scenes and removes downtime. This could use some implementation of technology analogous to Story Segmentation, which was a task for TRECVID 2003 [2], or using representative key frame extraction. However, for anything like Story Segmentation or key frame extraction to be possible, it is essential that the system first defines the video's scenes. To do this, we must find where scenes end, hence we use various shot boundary detection methods.

3.2   Cut Detection and Video Game Live Streaming

There has been siginificant research in the field of shot boundary detection over the past couple of decades, with multiple surveys of traditional shot detection methods [4] [6]. The shot boundary detection is a fundamental part of the broader field of video segmentation, which refers to the partitioning of video into spatial, temporal, or spatiotemporal regions that are homogeneous in some feature space. Video segmentation itself is an essential part of many video analysis problems, including, video summarization, indexing, retrieval, coding, authoring, and editing, as well as other applications in video surveillance and 3D motion and structure estimation with multiple moving objects [1]. There are many different aspects of the shot boundary detection problem, due to the fact that shot boundary's are can be defined by multiple different types of scene transitions - either an abrupt cut, or a gradual transition, such as a fade or a dissolve. These different transitions have different properties and patterns (see figure 1) that the shot boundary detection system must analyze, usually with each type of transition analyzed independently. For the purposes of this project, we will implement systems which detect abrupt cuts in video, and then apply these to systems to video game live stream clips.
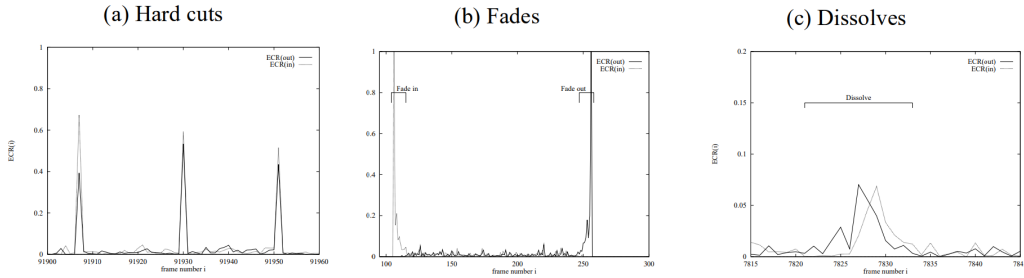

Figure 1: Gradual Transitions: Their Differing ECR Values [4]

Additionally, there has been as far as I know at the time of this project, no research on the topic of video game live stream cut detection. It is understandable why - the video game live streaming industry is still very young, and the change in domain when compared to more typical edited video or film is significant. Rather than there being a typical film or video editor defining the end of each scene, we must analyze and formulate the typical "scenes" that define contiguous activity in video game live streams. This

project will examine the different possible approaches to defining scenes in video game live streams, and attempt to formulate a schema for which we can apply traditional cut detection techniques, as well as outline areas of potential new developments for specifically this area, similar to what has been done for sports television video segmentation.

4 Methodology

4.1 What is Shot Boundary Detection?

While the number of possible types of scene transitions is quite large, in practice, the vast majority of all transitions are either hard cuts, fades, or dissolves.

5 Results

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

References

[1] A.C. Bovik. *The Essential Guide to Video Processing.* Elsevier Science, 2009.

[2] Lekha Chaisorn, Tat-Seng Chua, Chun-Keat Koh, Yunlong Zhao, Huaxin Xu, Huamin Feng, and Qi Tian. A two-level multi-modal approach for story segmentation of large news video corpus. 01 2003.

[3] Mansoor Iqbal. "Twitch Revenue and Usage Statistics (2020)". *Business of Apps*, Mar 2020.

[4] Rainer W Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 290–301. International Society for Optics and Photonics, 1998.

[5] Nick Statt. "YouTube is a $15 billion-a-year business, Google reveals for the first time". *The Verge*, Feb 2020.

[6] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(2):168–186, 2007.