Digital game video summarization based on screen and player videos

Natália Tiemi Yada
Coord. de Engenharia de Software
UTFPR
Dois Vizinhos, Paraná, Brazil
natalia.yada@gmail.com

Jhonatan Souza
Departamento de Informática
UFPR
Curitiba, Paraná, Brazil
jhonatansouza@ufpr.br

André Roberto Ortoncelli
Coord. de Engenharia de Software
UTFPR
Dois Vizinhos, Paraná, Brazil
ortoncelli@utfpr.edu.br

***Abstract*—This work presents an approach to summarize digital game videos based on motion, texture of videos of screen and player action. Two cameras it was used to collect videos of players: one directed to the face and other directed to the hands. Videos from four game match of two platform games were used to validate the proposed approach. An algorithm based on the maximization of a cost function, was used to select a predefined number of key frames - the video summary. In the experiments, all videos were summarized with different combinations of features/parameterizations. The best summaries were selected by the opinion of 23 people. Participants selected as the best results the summaries based on the videos of player action.**

***Index Terms*—video summarization, game videos, computer vision**

I. Introduction

In the last few years, it is possible to observe a popularization of game video streaming platforms, such as Twitch.tv and Youtube Live [1]. These platforms allow the online distribution of user-generated live video, let the live-streaming of video games a recent phenomenon [2].

Due to the volume of game videos available, is becoming increasingly important, the development of methods to efficiently browse, manage, and retrieve these videos.

In this context, there are in the literature, methods that focus on the streaming architecture improvements [3], [4], and also works related to the game video summarization, whose goal is to produce yet comprehensive summary to enable an efficient browsing experience [5]–[8].

In the context of summarization, the works have focused efforts on the production of techniques to specific games: Angry Birds [5] and League of Legends (LoL) [6]–[8].

In [5] a method called TGIF based on optical flow was proposed to detect TNT explosions in the game Angry Birds. This work is based on the hypothesis that explosions promote more emotions.

Visual features (motion and color), the event features (time and number of players), and viewer's behavior (number of viewers and number of emotion symbols in the Twitch), it was explored in [6] to detect highlights of the LoL game with a Support Vector Machine model.

A Deep Learning method to produce LoL highlights was proposed in [7], [8]. This method is based on the streamer's emotional state and behavior via webcam footage and audio.

Despite efforts to produce summaries published in the literature, these works focus on a limited set of games. There is no consensus about the best summarization approach or about the perfect set of characteristics that should be explored. In this context, there is space for more researches related to the game video summarization.

In this context, this work presents a game video summarization method that explores data about player's videos and screen's video. These videos were collected with the experimental configuration shown in Fig. 1.

Our work is the first approach that explores videos of the player's face and player's interaction with the keyboard/mouse to produce summaries. We extract color, motion, and texture from these videos. This set of information is the parameters of an algorithm based on the maximization of a cost function. The proposed algorithm selects a predefined number of key frames of the screen's video.

To validate the proposed approach, experiments were conducted with videos of two games played by two players. The following games it was used: i) Super Mario Bros - Star Scramble 3[1] and ii) Mario Jumping Star 2[2]. For each one of the experiments, the proposed algorithm was executed with different parameterizations.

The best summaries (the set of key frames that better represent the game video) were selected by the opinion of 23 people. Participants selected as the best result the summaries based on the players' videos.

The remaining of this article is organized as follows. Section II describes the proposed summary approach. Section III has details about the experiments. Section IV presents the analysis of the results. Section V concludes the paper and present future works.

II. Proposed Approach

The proposed video summarization approach uses three videos composed of n frames (V_s, V_f and V_h). V_s =

[1]https://www.clickjogos.com.br/jogos/super-mario-bros-star-scramble-3/
[2]https://www.youtube.com/watch?v=8xNZhkzWabA

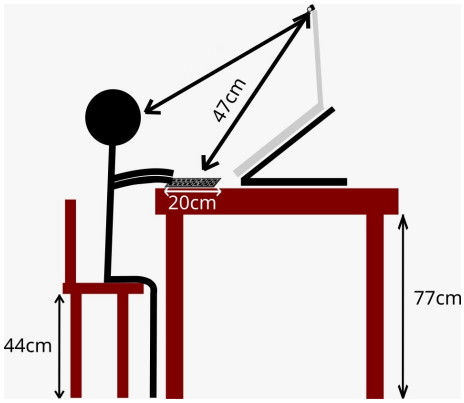

Fig. 1. Experimental environment.

$\{is_1, ..., is_n\}$ represents the game screen, $V_f = \{if_1, ..., if_n\}$ represents the player's face and $V_h = \{ih_1, ..., ih_n\}$ that represents the interaction of the player's hands with the keyboard/mouse. The experimental environment used to to collect the videos is in Fig. 1.

The value of n can be different for each set of the videos. Each set of videos it was manually synchronized, then the frames $is_X \in V_s$, $if_X \in V_f$ and $ih_X \in V_h$, it was related to the same instant of time. Can occurs millisecond variations in the frame synchronization due to the camera's characteristics.

The videos V_s, V_f and V_h it was used to create a set of m key frames, denoted as $K = \{ik_1, ..., ik_m\}$. Let $m < n$. In our experiments we work with $m = 10$.

To select the key frames, the our algorithm selects m frames based on the maximization of a cost function ψ (1), which calculates the difference of color ($g(V, x)$), texture ($l(V, x)$), motion ($m(V, x)$) and player's face size ($f(V, x)$).

Equation (1) receive a set of fourteen parameters $[x, V_g, V_r, V_m, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}]$. The x value represents one frame of the videos V_s, V_f and V_h. The last ten parameters (α_x values) it was used to defines the weight of the equation's terms. Each α_x parameter is a binary value, which indicates whether or not a given term will be used to compute the value.

$$\begin{aligned}\psi(x, V_g, V_r, V_m, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}) = \\ g(V_s, x) * \alpha_1 + g(V_h, x) * \alpha_2 + g(V_f, x) * \alpha_3 \\ + l(V_s, x) * \alpha_4 + l(V_h, x) * \alpha_5 + l(V_f, x) * \alpha_6 \\ + m(V_s, x) * \alpha_7 + m(V_h, x) * \alpha_8 + m(V_f, x) * \alpha_9 \\ + f(V_f, x) * \alpha_{10}\end{aligned} \quad (1)$$

The color information is calculated with the $g(V, x)$ function, which returns the Euclidean distance between the gray level intensity histogram of the frames: $i_x, i_{x+1} \in V$.

The $l(V, x)$ function defines the texture difference following two steps: i) the texture of the frames i_x, i_{x+1} inV is described with the Local Binary Pattern (LBP) operator [9]; and ii) the Euclidean distance between the texture images is returned.

The motion information was computed with the function $m(V, x)$, which returns the sum of the subtraction of each pair of adjacent pixels in the frames $i_x, i_{x+1} \in V$, considering gray-level images.

The function $f(V, x)$ returns the difference between the size of the player's face in the frames $i_x, i_{x+1} \in V$. To identify the faces, we used the OpenCV library [10].

We use the face size difference because we believe that zoom in/zoom out of the face may be related to the player's attention. We believe that the face zoom in occurs when the game is more difficult or more interesting.

The values of the $g(V, x)$, $l(V, x)$, $m(V, x)$ and $f(V, x)$ are normalized between 0 and 1, so the value obtained with (1) can vary between 0 and 10, depending on the α parameters.

III. Experiments

To analyze the prioritization approach presented in this work, a database was created with videos recorded of two players playing two different platform games. Each of the players was recorded playing each of the two selected games, so the experimental database is composed of four sets of videos.

The selected games were: Game 1 - Super Mario Bros - Star Scramble; Game 2 - Mario Jumping Star. Game 1 had previously been played by the players, and game 2 was not known by the players.

Both the selected games are based on the Mario Bross classical game, with similar rules and layout. Game 2 differs from the classical Mario Bross by the design of the game scene components. Game 1 is played with the keyboard, and Game 2 is played with the mouse.

An ID identifies the videos in our database. Table I show the video ids and the respective game/player. For each of the four sets of videos, the proposed summarization algorithm was applied with seven different parameter configurations. Table II as the parameters configuration.

Fig. 2, as the summaries computed in the experiments carried out with the videos with id 2, respectively. These figures are organized in seven lines that correspond to a set of key frames selected with each parameterization of the Algorithm 1 (Table 2).

TABLE I
Experimental Configuration - Id, video and player

Videos ID	Game	Player
1	1	1
2	2	1
3	1	2
4	2	2

To evaluate the experiments, 23 people were recruited to select the best summary produced for each video set. Each person watched the videos and then selected the best summary for each set of videos. The experimental results, as in Table III, that was the number of people that were selected each summary.

Fig. 2. Key frames of the videos with id 2.

TABLE II
EXPERIMENTAL CONFIGURATION - α PARAMETERS

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
C1	0	0	1	0	0	1	0	0	1	0
C2	0	1	0	0	1	0	0	1	0	1
C3	1	0	1	0	0	1	0	0	0	0
C4	0	0	0	1	1	1	0	0	0	0
C5	1	1	1	0	0	0	0	0	0	0
C6	0	0	0	0	0	0	1	1	1	0
C7	0	1	1	0	1	1	0	1	1	1

TABLE III
EXPERIMENTAL RESULTS

Experiment Id	C1	C2	C3	C4	C5	C6	C7
1	**8**	2	2	3	3	0	5
2	**9**	3	4	0	0	2	5
3	0	**15**	3	0	2	3	0
4	0	**10**	3	3	4	3	0

IV. ANALYSIS OF RESULTS

In the experiments carried out with the videos with id 1 and 2, the best summarization results were obtained with the C1 configuration, which obtained 34.8% of the votes for the video id 1 and 39.1% of the votes for the video with id 2. The C1 configuration explores only information from the vídeo of the player's face.

In the experiments carried out with the videos with id 3 and 4, the best summarization results were obtained with the C2 configuration, which obtained 65.2% and 45.5% of the votes, respectively. The C2 configuration explores only information from the video the player interaction with the keyboard/mouse.

It should be noted that in all experiments, on average, 60.7% of the participants chose summaries produced solely with information from the players' face and hands (experimental configurations C1, C2, and C7).

V. CONCLUSION

This work presents an approach to summarize digital game videos based on motion, the texture of player's and videos screen's videos. The development of this type of approach is relevant because with the popularization of the game video streaming platform, it is becoming increasingly important the development of methods to browse, manage, and retrieve these videos efficiently.

Experiments were carried out based on the opinion of 23 people. Most people chose the summaries produced with information obtained from the player's videos (face and hands videos).

Although the results are still inconclusive (due to the small volume of games and experiments carried out), these results indicate that explore player information to summarize digital games videos is a promising approach.

In future works, we intend to expand our database with videos of different games. We also intend to explore a larger set of information to produce summaries, such as facial expressions and emotions.

In future works, in addition to videos of player interaction with the keyboard/mouse, can be explored different forms of interaction with the game, such as joystick and tangible interfaces.

REFERENCES

[1] E. Harpstead, J. S. Rios, J. Seering, and J. Hammer, "Toward a twitch research toolkit: A systematic review of approaches to research on game streaming," in *Annual Symposium on Computer-Human Interaction in Play*, 2019, pp. 111–119.

[2] P. Lessel, A. Vielhauer, and A. Krüger, "Expanding video game live-streams with enhanced communication channels: A case study," in *onference on Human Factors in Computing Systems*, 2017, pp. 1571–1576.

[3] D. C. Barboza, D. C. Muchaluat-Saad, E. W. G. Clua, and D. G. Passos, "An architecture for 2d game streaming using multi-layer object coding," in *Brazilian Symposium on Computer Games and Digital Entertainment*. IEEE, 2019, pp. 62–71.

[4] J. Sun and M. Claypool, "Evaluating streaming and latency compensation in a cloud-based game," in *Advanced International Conference on Telecommunications*, 2019.

[5] C. Yang, P. Paliyawan, R. Thawonmas, and T. Harada, "Tgif!: Selecting the most healing tnt by optical flow." in *AAAI Spring Symposium: Interpretable AI for Well-being*, 2019.

[6] W.-T. Chu and Y.-C. Chou, "On broadcasted game video analysis: event detection, highlight detection, and highlight forecast," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9735–9758, 2017.

[7] C. Ringer and M. A. Nicolaou, "Deep unsupervised multi-view detection of video game stream highlights," in *Proceedings of the 13th International Conference on the Foundations of Digital Games*, 2018, pp. 1–6.

[8] C. Ringer, J. A. Walker, and M. A. Nicolaou, "Multimodal joint emotion and game context recognition in league of legends livestreams," in *IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.

[9] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer vision using local binary patterns*. Springer Science & Business Media, 2011, vol. 40.

[10] K. Goyal, K. Agarwal, and R. Kumar, "Face detection and tracking: Using opencv," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1. IEEE, 2017, pp. 474–478.