Gameplay genre video classification by using mid-level video representation

Renato Augusto de Souza[‡], Raquel Pereira de Almeida[‡], Arghir-Nicolae Moldovan[*],
Zenilton Kleber G. do Patrocínio Jr.[‡], Silvio Jamil F. Guimarães[‡]

[‡]Audio-Visual Information Proc. Lab. (VIPLAB)
Computer Science Department – ICEI – PUC Minas
{renato.souza,raquel.almeida,zenilton,sjamil}@pucminas.br
[*]School of Computing, National College of Ireland, Dublin, Ireland
arghir.moldovan@ncirl.ie

***Abstract*—As video gameplay recording and streaming is becoming very popular on the Internet, there is an increasing need for automatic classification solutions to help service providers with indexing the huge amount of content and users with finding relevant content. The automatic classification of gameplay videos into specific genres is not a trivial task due to their high content diversity. This paper address the problem of classifying video gameplay recordings into different genres by using mid-level video representation based on the BossaNova descriptor. The paper also proposes a public dataset called GameGenre containing 700 gameplay videos groped into 7 genres. The results from experimental testing show up to 89% classification accuracy when the gameplay videos are described by BossaNova descriptor using BinBoost as low-level image descriptor.**

***Keywords*-Gameplay videos; gameplay genre video classification; mid-level video representation, BossaNova video descriptor.**

I. Introduction

Gameplay is the specific way in which players interact with a game and in particular with a video game [1]. Recently there has been a tremendous increase in user-generated video gameplay recording and streaming on dedicated online platforms such as Twitch and YouTube Gaming [2]. In 2015, Twitch had over 1.7 million broadcasters streaming every month and over 120 million monthly viewers [3], while the top 100 YouTube gaming channels generated over 7.2 billion views in January 2016 [4]. Given the huge amount of video content in general and gameplay videos in particular there is an increasing need for automatic video processing and analysis techniques such as event detection, indexing and classification, in order to help the service providers with managing the content and the users with finding relevant videos [5].

In this context, we address the problem of automatic classification into different genres of gameplay videos that were recorded for distribution on the Internet. While much research has been conducted on video classification in general [6], [7], to the best of our knowledge, this is the first work about gameplay videos classification. This paper also proposes a public dataset of gameplay videos, with the particular purpose to evaluate the proposed classification strategies. The dataset named GameGenre, consists of 700 videos (more than 116 hours), classified into 7 game genres.

Previous research works in the area have mainly focused on classifying video content into generic categories such as news, sports or movies [6], [7]. Significant research effort was also made on further sub-classifying specific video genres, such as movies into action, comedy, horror, etc. [8], [9], or sports into soccer, tennis, volleyball, etc. [10]. Other research works have focused on more specific areas such as classifying educational videos [11], pornography video classification [12], [13], game classification in order to identify gameplay bricks [14], or large scale classification of sport videos [15], among others.

The task of video classification is usually based on: (i) low-level feature extraction, (ii) mid-level description generation, and (iii) video classification. Feature extraction is the main vision task in classification pipeline and consists in extracting visual information from video. The mid-level descriptions are representations that aggregate all video information into only one descriptor, usually a Bag-of-Visual-Words. Finally, the video classification step involves activities of learning statistical models based on mid-level descriptors, and applying those models to classify new observations.

In a typical approach to video classification, the stage of feature aggregation is performed on all extracted features, however this strategy may suffer from two issues: polysemy and synonymy. To cope with these problems, instead of computing a traditional Bag-of-Visual-Words, the video descriptors are based on a recent mid-level representation, called BossaNova. In this work, we propose to classify the gameplay genre videos by using a BossaNova Video Descriptor, which was proposed for pornography classification [12], [13], and a Bag-of-BossaNova descriptors.

The main contributions of this research work are twofold: (i) proposal of a new public video dataset, and (ii) study of new mid-level video descriptions applied to gameplay genre video classification.

This rest of this paper is organized as follows. Section II and Section III describe some image and video representations. Section IV presents the public dataset GameGenre with some examples. Section V shows a quantitative analysis of the proposed approach for gameplay genre video classification. Finally, in Section VI, concludes the paper and presents some further work directions.

II. Image Representations

Both the effectiveness and the efficiency of an image or video processing system are dependent on *descriptors* (or visual features). A feature extraction algorithm can produce either a single feature vector or a set of feature vectors. In the former case, a single feature vector must capture the entire information of the visual content (named *global descriptor*). In the latter case, a set of feature vectors (or *local descriptors*) is associated with the image visual content.

Thanks to the development of very discriminant low-level local features (such as SIFT descriptors [16]), and the emergence of mid-level aggregate representations, based on the quantization of those features (such as the Bag-of-Words[1] model [17]), significant progress has been made in visual recognition tasks. In order to be able to efficiently deal with a large number of local descriptors, an important task is the construction of a visual dictionary, or codebook. Afterwards, the codebook can be used to create a *mid-level* image descriptor – named *Bag-of-Words* (BoW) – to describe any image using two steps: *coding* (*i.e.*, assignment of descriptors to visual words), and *pooling* (*i.e.*, generation of an image (or video) representation). The BoW is simple to build, however it may suffer from two issues: *polysemy* – a single visual word may represent different contents; and *synonymy* – several visual words may characterize the same content.

In general, for the video classification task, the video must be described by a mid-level representation (or a global descriptor). The following section reviews some key concepts involving low-level descriptors and mid-level image representation.

A. Low-level Features

A low-level feature descriptor can be considered as a function applied to a region of the image to perform its description. The simplest way to describe a region is to represent all the pixels in this region in a single vector. However, depending on the information to be described, this would result in a high-dimensional vector leading also to a high computational complexity for a future recognition of this region [18]. In this section, we review two low-level descriptors, which can be classified in two distinct ways [19]: (i) non-binary descriptors, and (ii) binary descriptors.

SIFT – Scale Invariant Feature Transform: One of the most important descriptors used in the literature is the SIFT [16]. This descriptor performs a scale-space analysis leading to a great performance according to the scale invariance. Although the author has developed the SIFT descriptor to be used on object recognition tasks, it has become the most widely used descriptor in several other applications. This is due to its high discriminative power and stability.

To describe each patch, an orientation α is assigned selecting the angle that represents the histogram of local gradients (calculated for each pixel around the keypoint). Then, the region of points around the keypoint, oriented by α, is divided into subregions composed by a grid of size $G \times G$. Next, a histogram of orientation consisting of B bins is created from the samples of each subregion. The descriptor is then obtained from the concatenation of the histograms of these subregions, composed of $G \times G \times B$ values. The default values for G and B are usually 4 and 8, respectively, resulting in a vector of 128 length. Finally, the descriptor is normalized turning it robust to illumination variations.

BinBoost: Trzcinski et al. [20] proposed a new framework with the aim of creating a binary descriptor extremely compact and highly discriminative. The BinBoost descriptor is robust to changes in lighting and viewpoint. Each bit generated by BinBoost is computed by using a binary hash function, the same way as the AdaBoost classifier does [21]. This function is based on weak learners that take into account the orientation intensity of gradients on the patch to be described. The hash function is optimized iteratively, i.e., at each iteration, incorrect samples are assigned to a greater weight while the weight of the correct samples is decreased. In this way, the next bits to be calculated will tend to correct the error of their predecessors.

B. Mid-level Image Representation

The BossaNova is a mid-level image representation [22], which offers more information-preserving pooling operation based on a distance-to-codeword distribution. The BossaNova approach follows the BoW formalism (coding/pooling), but it proposes an image representation which keeps more information than BoW during the pooling step, since it estimates the distribution of the descriptors around each codeword, by computing a histogram of distances between the descriptors found in the image and those in the codebook. In BossaNova code, the authors proposed a soft-assignment strategy considering only the k-nearest codewords for coding a local descriptor.

This representation was successfully applied to the context of visual recognition. In comparison to BoW, BossaNova significantly outperforms it. Furthermore, by using a simple histogram of distances to capture the relevant information, the method remains very flexible and keeps the representation compact. For those reasons, we choose the BossaNova approach as the mid-level feature to be used in the experiments.

III. Video Representations

Some applications for video classification are based on majority voting [23], however this method works on binary classification. The gameplay genre video is a multiclass classification which makes prohibitive the majority voting. In this sense, we propose the use of two mid-level video representations which aggregates mid-level image representation.

[1] Bag-of-Words models have blurred somewhat the distinction between local and global descriptors, because they propose a single (global) feature vector based upon several (local) features.

A. *Bag-of-BossaNova Descriptor*

The Bag-of-BossaNova descriptor is a Bag-of-Visual-Words in which the features are BossaNova descriptors computed for each video frame. Thus, each video is represented by one Bag-of-BossaNova. It is important to note that a BossaNova representation aggregates any kind of low-level features. In this work, we have considered both SIFT and BinBoost in order to compute it.

B. *BossaNova Video Descriptor*

The BossaNova Video Descriptor (BNVD) was proposed in [12], [13] for pornography recognition and it is based on the combination of mid-level representations. In fact, this descriptor can be considered as a simple strategy that aggregates information of the mid-level representations of all video frames into a single representation by using an operator, such as *median*, *max* or *min*. The procedure is explained in the following.

Let $\mathscr{V}$ be a video sequence. $\mathscr{V} = \{f^i\}$, $i \in [1,N]$, where f^i is the keyframe[2] of the shot i and N is the number of keyframes. Let $\mathscr{X} = \{\mathbf{z}^i\}$, $i \in [1,N]$ be a set of BossaNova vectors computed for the video $\mathscr{V}$ in which $\mathbf{z}^i$ is a BossaNova vector extracted for the keyframe f^i. Let $\mathbf{O}$ and $\mathbf{P}$ be two functions for aggregating the information of BossaNova and the Bag of Visual Words. The BossaNova Video Descriptor (BNVD) can be modeled by a function $\mathbf{W}$ as follows:

$$\begin{aligned}
\mathbf{O} : \mathbb{R}^B &\longrightarrow \mathbb{R}^B, \\
\mathbf{P} : \mathbb{R}^M &\longrightarrow \mathbb{R}^M, \\
\mathbf{W} : \mathbb{R}^Z &\longrightarrow \mathbb{R}^Z, \\
\mathscr{X} &\longrightarrow \mathbf{W}(\{\mathbf{z}^i\}) = \left[\left[o_{m,b}\right], p_m\right]^{\mathrm{T}}, \\
o_{m,b} &= \mathbf{O}(\{z^i_{m,b}\}), \\
p_m &= \mathbf{P}(\{t^i_m\}), \qquad (1)
\end{aligned}$$

where $Z \subset \{1,...,M\} \times \{1,...,B\}$, and $\mathbf{z}^i = \left[\left[z^i_{m,b}\right], t^i_m\right]^{\mathrm{T}}$.

Intuitively, this new video descriptor represents a relation for each codeword to the codebook, since each BossaNova representation contains information regarding the distance-to-codeword distribution. The main goal of applying the functions $\mathbf{O}$ and $\mathbf{P}$ to the BossaNova vectors, in order to compute $o_{m,b}$ and p_m, respectively, is to employ a filtered-like operation to the entire video content which is represented by this mid-level representation.

As discussed in [12], [13], this descriptor, intuitively, represents the median distance for each visual word to the codeword, since each BossaNova representation contains information about the distance-to-codeword distribution. Moreover, outliers are eliminated by the *median* function. During experiments, we have also used the *max* operator to aggregate the information of video frames. In this case, we expect that BNVD can represent more saliently the video information.

TABLE I
GAMEGENRE DATASET SUMMARY.

Videogame Genre	**Total Duration**
Card	18hr 38min 04 sec
Fighting	15hr 54min 34sec
First Person Shooter	16hr 31min 59sec
Racing	15hr 49min 41sec
Real Time Strategic	15hr 28min 59sec
Side-Scrolling	16hr 54min 0sec
Team Sport	16hr 53min 15sec
Total	**116hr 10min 32sec**

IV. THE GAMEGENRE DATASET

The proposed public GameGenre dataset[3] is divided into 7 classes representing different genres of gameplay videos. Each class contains 100 videos downloaded from YouTube, totalling 700 videos in the dataset. The videos were selected aiming for content variety within each class, are encoded at high resolution and the duration of each video varies from 4 minutes up to 20 minutes. Table I presents a summary of the dataset size in terms of total videos duration for each gameplay video class.

A. *Genre Definitions*

Defining a game genre is a very challenging and ambiguous task [24], since a game could fit into various genres. A title that exemplifies well this scenario is Grand Thief Auto (GTA) whose genre is defined by its creators as action-adventure, third person shooter and racing. Adding to that, the game may be hybrid and use different technologies and platforms.

As a way to avoid these difficulties we used the genre characterization proposed in [25] to specify the genres, combined with some aspects proposed in [26] to separate the dataset into different classes. The selected aspects were Gameplay, Temporal and Style, in which Gameplay defines the experience of the player, Temporal defines the relation with time (e.g., real time or time manipulation), while Style defines the model of the game, such as shooter or action. Following we describe each genre present in the GameGenre dataset.

Real Time Strategic (RTS): A strategic game consists in challenges that players should choose from a wide range of possibilities, actions and movements to complete a mission or objective. An RTS game can be seen as a subclass of strategic game, but even more challenging since it demands that multiple players interact simultaneously. Fig. 1, presents two examples of videos from the RTS class.

Card: Card games try to simulate real life card games, such as poker and imaginary, in which a player can choose to play with other players or with the machine. It can also be seen as subclass of strategic games, since the player needs to plan his actions, gather data during the game and consider its options before each movement. Fig. 2, presents two examples of videos from this class.

[2] A keyframe is a frame that represents the content of a logical unit, like a shot or scene, for example.

[3] Available at http://www.icei.pucminas.br/projetos/viplab/databases/gamegenre

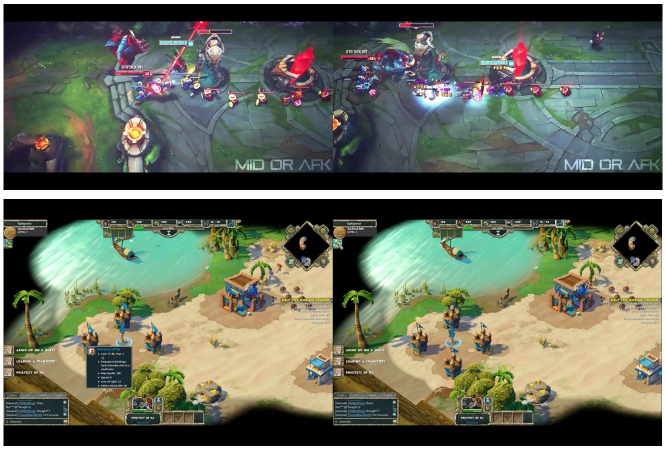

Fig. 1. Example of video class *Real Time Strategic*: (top) League of Legends; and (bottom) Age of Empires.

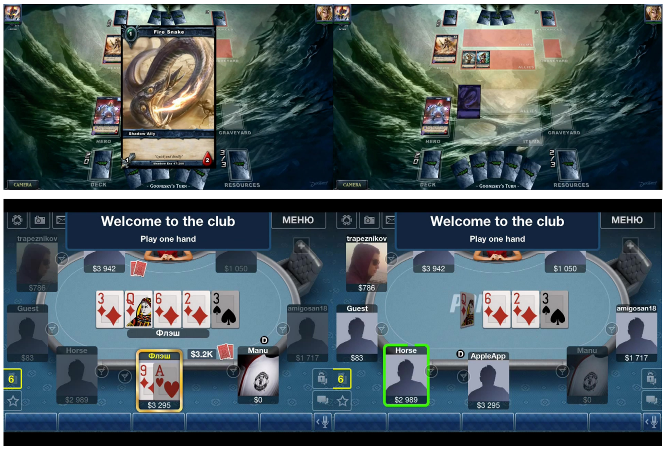

Fig. 2. Example of video class *Card*: (top) Shandow Era; and (bottom) Texas Poker.

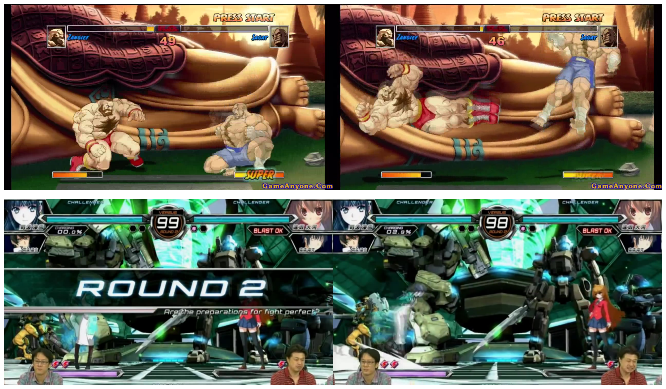

Fig. 3. Example of video class *Fighting*: (top) Super Street Fighter; and (bottom) Dengeki Bunko.

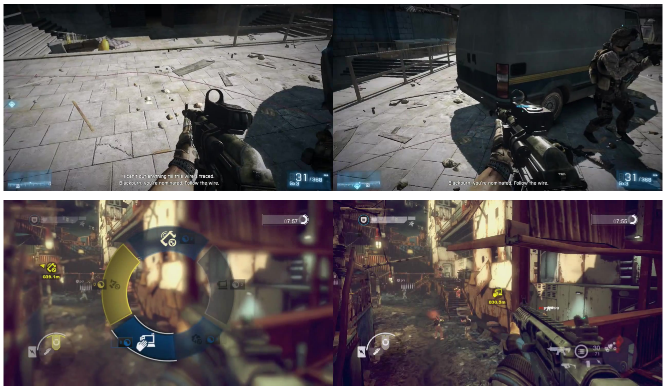

Fig. 4. Example of video class *First Person Shooter*: (top) Battle Field 3; and (bottom) Brink the Objectives.

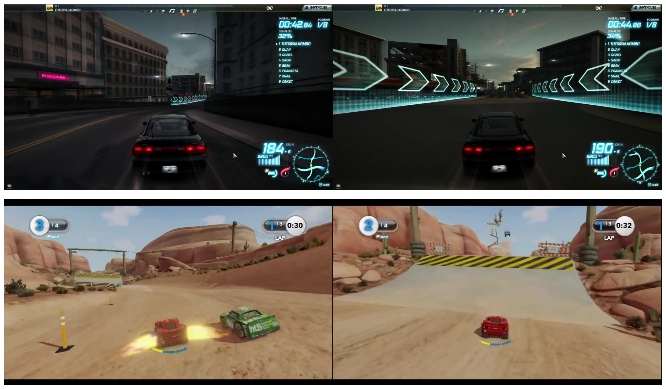

Fig. 5. Example of video class *Racing*: (top) Need for Speed; and (bottom) Car Alive.

Fighting (Fight): Adams [27] declares that fighting games require physical skills, such as reaction reflex and timing. For these aspects it can be seen as a subclass of action games. The game mechanics consist of combats using fighting techniques. The player usually combats another player and in certain cases, with the machine. Fig. 3, presents two examples of videos from this class.

First Person Shooter (FPS): The focus of this genre of game is the vision of the protagonist. The game consists in shooting at specific points, like enemies or strategical spots, and its difficulty increases according to the the levels achieved by the player. Fig. 4, presents two examples of videos from this class.

Racing (Race): The main purpose of this genre is to simulate the action of driving or piloting a real vehicle. In general, the player's objective consists on doing some quest faster than other players or the machine. Fig. 5, presents two examples of videos from this class.

Side-Scrolling (Side): Present in several of earliest video game platforms, side-scrolling or platforms consist in an avatar moving forward and backward with limited abilities, collecting items and confronting enemies. Fig. 6, presents two examples of videos from this class.

Team Sport (Team): Sport games focus on simulating real sport matches. As the name implies, Team Sport games has the purpose to simulate a sport with more than one character, such as soccer, volleyball and rugby. Fig. 7, presents two examples of videos from this class.

V. Experiment

In this section, we describe the experimental setup and discuss the results obtained by applying the BossaNova video

Fig. 6. Example of video class *Side-Scrolling*: (top) BloodRayne Betrayal; and (bottom) BroForce.

Fig. 7. Example of video class *Team Sport*: (top) NCAA Basketball 10; and (bottom) College Lacrosse.

descriptor for gameplay genre video classification.

A. Experimental Setup

In order to compute the mid-level video representations, we extracted one frame by minute followed by a feature extraction using two descriptors: SIFT and BinBoost. For the SIFT descriptor, we adopted the code from OpenCV's repository [28]; while for BinBoost we have used the code publicly available[4]. The code for BossaNova is also publicly available[5].

Two different representations were used in order to compute a global signature for videos: the BossaNova Video Descriptor and the Bag-of-BossaNova. For each approach we used two different low-level features: SIFT and BinBoost with dense sample in three different window size (6, 20 and 30 pixels). We kept the BossaNova parameter values the same as in [23], [12] ($B = 2, \lambda_{min} = 0.4, \lambda_{max} = 2.0, s = 10^{-3}$), and we used two different aggregating functions: *max* and *median*. For creating the codebooks, we have used 128 centroids. For classification we used non-linear SVM with an RBF kernel. For parameter definition during training phase, we used a grid search with cross-validation. In scaled tests we normalized the values between 0 and 1.

B. Quantitative Analysis

Our experiments are based on 5-fold cross validation strategy. For computing the accuracy (presented in Table II), we have computed the average value for all classifications.

TABLE II
ACCURACIES FOR THE GAMEPLAY GENRE VIDEO CLASSIFICATION USING TWO DIFFERENT MID-LEVEL VIDEO REPRESENTATIONS AND TWO DIFFERENT LOW-LEVEL IMAGE DESCRIPTORS.

Approach	Feature	Window size	Accuracy Normalization no	Accuracy Normalization yes
Bag-of-BossaNova	BinBoost	30	**89.84%**	**89.51%**
	SIFT	6	83.06%	83.96%
	SIFT	20	83.47%	82.82%
	SIFT	30	85.63%	84.65%
Bossa Nova VD (max)	BinBoost	30	79.31%	85.18%
	SIFT	6	80.16%	80.94%
	SIFT	20	80.53%	80.65%
	SIFT	30	78.41%	78.73%
Bossa Nova VD (median)	BinBoost	30	86.86%	87.27%
	SIFT	6	75.59%	82.94%
	SIFT	20	83.88%	75.96%
	SIFT	30	81.67%	76.12%

From our experiments, we observe a very interesting behaviour. The BossaNova Video descriptor in conjunction with BinBoost is much better than BossaNova Video descriptor in conjunction with SIFT. Furthermore, as one can see, the Bag-of-BossaNova, in conjunction with BinBoost outperforms all tested strategies (with 89.84% of accuracy). We also applied a normalization on the data in order to minimize the noise, and as one can see, there is no impact of this operation on both mid-level video representations.

Table III, Table IV and Table V present the confusion matrices, for the test case with the highest accuracy result, in each of the three cases: using Bag-of-BossaNova representation, and using BossaNova Video Descriptor with max and median aggregation.

VI. CONCLUSIONS AND FURTHER WORKS

Gameplay is the specific way in which players interact with a game and in particular with video game. In this paper, we propose both a public dataset GameGenre for gameplay videos categorized by genres, as well as a methodology for classifying these videos by using mid-level video representation.

Two different mid-level video representations were studied, namely the Bag-of-BossaNova and BossaNova Video descriptor, and for each case the video frames were described by using bothe SIFT and BinBoost. With the BossaNova flexibility and ability to capture the relevant information extracted with the low-level feature descriptors, combined with Bow to aggregate the information obtained with the *mid-level* step, we were able to obtain really good results. The experimental results have shown that Bag-of-BossaNova descriptor with BinBoost achieves 89.84% classification accuracy, outperforming all other tested methods.

[4] http://www.cvlab.epfl.ch/research/detect/binboost
[5] http://www.npdi.dcc.ufmg.br/bossanova

TABLE III

CONFUSION MATRIX FOR THE CLASSIFICATION WITH THE BEST ACCURACY (91.22%) WHEN USING BAG-OF-BOSSANOVA REPRESENTATION.

	Card	Fight	FPS	Race	RTS	Side	Team
Card	50%	0%	5%	10%	10%	15%	10%
Fight	5%	80%	0%	5%	5%	0%	5%
FPS	0%	0%	95%	0%	0%	0%	5%
Race	5%	20%	0%	75%	0%	0%	0%
RTS	10%	10%	0%	0%	50%	20%	10%
Side	25%	0%	0%	0%	5%	65%	5%
Team	0%	0%	10%	10%	5%	5%	70%

TABLE IV

CONFUSION MATRIX FOR THE CLASSIFICATION WITH THE BEST ACCURACY (86.73%) USING BOSSA NOVA VIDEO DESCRIPTOR REPRESENTATION WITH *max* OPERATOR FOR AGGREGATING THE IMAGE DESCRIPTORS.

	Card	Fight	FPS	Race	RTS	Side	Team
Card	50%	0%	5%	5%	10%	20%	10%
Fight	20%	30%	10%	10%	5%	10%	15%
FPS	0%	5%	70%	5%	15%	5%	0%
Race	10%	10%	10%	55%	5%	5%	5%
RTS	5%	15%	0%	0%	75%	5%	0%
Side	0%	15%	40%	0%	15%	30%	0%
Team	15%	0%	0%	10%	5%	5%	65%

For further work, will study the behaviour of different low-level features and other video representations in gameplay genre video classification.

ACKNOWLEDGMENTS

The authors are grateful to PUC Minas, CAPES, FAPEMIG (PPM-00006-16) and CNPq for the financial support to this work.

REFERENCES

[1] C. A. Lindley, *Technologies for Interactive Digital Storytelling and Entertainment: Second International Conference, TIDSE 2004, Darmstadt, Germany, June 24-26, 2004. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ch. Narrative, Game Play, and Alternative Time Structures for Virtual Environments, pp. 183–194. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-27797-2_25

[2] K. Pires and G. Simon, "YouTube Live and Twitch: A Tour of User-generated Live Streaming Systems," in *Proceedings of the 6th ACM Multimedia Systems Conference*, ser. MMSys '15. New York, NY, USA: ACM, 2015, pp. 225–230.

TABLE V

CONFUSION MATRIX FOR THE CLASSIFICATION WITH THE BEST ACCURACY (88.57%) USING BOSSA NOVA VIDEO DESCRIPTOR REPRESENTATION WITH *median* OPERATOR FOR AGGREGATING THE IMAGE DESCRIPTORS.

	Card	Fight	FPS	Race	RTS	Side	Team
Card	70%	5%	5%	5%	0%	5%	10%
Fight	5%	45%	30%	0%	5%	0%	15%
FPS	5%	0%	75%	10%	10%	0%	0%
Race	0%	0%	25%	55%	10%	0%	10%
RTS	0%	5%	0%	0%	85%	5%	5%
Side	5%	10%	50%	0%	10%	25%	0%
Team	5%	0%	0%	10%	10%	10%	65%

[3] Twitch, "The 2015 Retrospective," 2016. [Online]. Available: https://www.twitch.tv/year/2015

[4] J. Cohen, "Top 100 Most Viewed YouTube Gaming Channels Worldwide," Mar. 2016. [Online]. Available: http://www.tubefilter.com/2016/03/30/top-100-most-viewed-youtube-gaming-channels-worldwide-february-2016/

[5] C. Tzelepis, Z. Ma, V. Mezaris, B. Ionescu, I. Kompatsiaris, G. Boato, N. Sebe, and S. Yan, "Event-based Media Processing and Analysis: A Survey of the Literature," *Image and Vision Computing*, 2016.

[6] D. Brezeale and D. Cook, "Automatic Video Classification: A Survey of the Literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 3, pp. 416–430, May 2008.

[7] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, 2011.

[8] Y.-F. Huang and S.-H. Wang, "Movie Genre Classification Using SVM with Audio and Video Features," in *Active Media Technology*, ser. Lecture Notes in Computer Science, R. Huang, A. A. Ghorbani, G. Pasi, T. Yamaguchi, N. Y. Yen, and B. Jin, Eds. Springer Berlin Heidelberg, Dec. 2012, no. 7669, pp. 1–10, dOI: 10.1007/978-3-642-35236-2_1.

[9] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie Genre Classification via Scene Categorization," in *Proc. ACM Multimedia 2010*. Firenze, Italy: ACM, Oct. 2010.

[10] F. Cricri, M. J. Roininen, J. Leppanen, S. Mate, I. D. D. Curcio, S. Uhlmann, and M. Gabbouj, "Sport Type Classification of Mobile Videos," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 917–932, Jun. 2014.

[11] A.-N. Moldovan, I. Ghergulescu, and C. H. Muntean, "Learning Assessment for Different Categories of Educational Multimedia Clips in a Mobile Learning Environment," in *Proceedings of 25th Society for Information Technology and Teacher Education International Conference (SITE 2014)*. Jacksonville, Florida, USA: AACE, Mar. 2014, pp. 1687–1692.

[12] C. Caetano, S. E. F. de Avila, S. J. F. Guimarães, and A. de Albuquerque Araújo, "Pornography detection using bossanova video descriptor," in *22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014*, 2014, pp. 1681–1685. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6952616

[13] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. GuimarÃ£es, and A. de A. AraÃºjo, "A mid-level video representation based on binary descriptors: A case study for pornography detection," *Neurocomputing*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216307111

[14] D. Djaouti, J. Alvarez, J.-P. Jessel, G. Methel, and P. Molinier, "A

gameplay definition through videogame classification," *Int. J. Comput. Games Technol.*, vol. 2008, pp. 4:1–4:7, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/470350

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, pp. 91–110, 2004.

[17] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1–8.

[18] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 10, pp. 1615–1630, 2005.

[19] A. Canclini, M. Cesana, R. A., M. Tagliasacchi, J. Ascenso, and C. R., "Evaluation of low-complexity visual feature detectors and descriptors," in *International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–7.

[20] V. L. T. Trzcinski, M. Christoudias and P. Fua, "Boosting Binary Keypoint Descriptors," in *CVPR*, 2013, pp. 2874–2881.

[21] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[22] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "Pooling in image representation: the visual codeword point of view," *CVIU*, vol. 117, no. 5, pp. 453–465, 2013.

[23] C. Caetano, S. Avila, S. Guimarães, and A. d. A. Araújo, "Representing local binary descriptors with bossanova for visual recognition," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, ser. SAC '14. New York, NY, USA: ACM, 2014, pp. 49–54. [Online]. Available: http://doi.acm.org/10.1145/2554850.2555058

[24] F. D. Laramée, *Game design perspectives*. Charles River Media, Inc., 2002.

[25] H. M. Chandler and R. Chandler, *Fundamentals of game development*. Jones & Bartlett Learning, 2011.

[26] J. H. Lee, N. Karlova, R. I. Clarke, K. Thornton, and A. Perti, "Facet analysis of video game genres," *iConference 2014 Proceedings*, 2014.

[27] E. Adams, *Fundamentals of game design*. Pearson Education, 2013.

[28] G. Bradski, "The {OpenCV} Library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120, 122–125, 2000.