DCU-ADAPT at MediaEval 2019: GameStory

Yasufumi Moriya, Gareth J. F. Jones
ADAPT Centre, School of Computing, Dublin City University, Ireland
yasufumi.moriya@adaptcentre,gareth.jones@dcu.ie

ABSTRACT

We describe the DCU-ADAPT participation in the GameStory task at MediaEval 2019. Our approach comprises of two stages: (i) finding replays in a commentator stream and (ii) comparing the found replays to player streams to identify source segments of the replays. To analyse event patterns of a commentator video, we built a neural network model to classify a video frame into three categories: game play, sponsor logo and audience or players. The classifier was applied to a commentator stream to obtain event patterns of the video. In the second stage, each of the found replay segments was compared against player streams using mean squared error (MSE) and structural similarity (SSIM) to identify source segments. On the test set, our approach found replays with precision 0.68, recall 0.5 and F1 0.57 when the Jaccard threshold was 0.5 and the average overlap of the replays with source segments was 0.31.

1 INTRODUCTION

E-sports has become popular entertainment for millions of people. The nature of e-sports data is multimodal, where a visual stream contains game plays, an audio stream contains commentaries and textual metadata contains various information about matches. GameStory at MediaEval 2019 offered a sub problem of summarisation of e-sports matches [3]. The goal of the task is to identify replays of a game play in a commentator stream and to locate the replays in player streams. Our approach to the task consists of two stages: (i) finding replays in a commentator stream based on visual event patterns and (ii) computing similarity between video frames from player streams and video frames from the found replays to locate sources of the replays.

Figure 1 shows our approach. To convert a commentator stream into event patterns, we built a neural network model that can classify a given video frame into one of the three categories: game play, sponsor logo and audience or players. After identifying replay segments, we applied optical character recognition (OCR) to the segments to identify match rounds. This reduces the search space to look for source segments of the replays. We computed a similarity score of each replay segment and player streams of a match round obtained using OCR. Replays found by our approach on the test set produced F1 0.57 when the Jaccard threshold was 0.5 and the average overlap of the replays with source segments was 0.31.

2 DATA DESCRIPTION

This section gives a brief of the provided task data, detailed information is contained in [3]. The videos were recorded in the tournament of the Counter-Strike: Global Offensive (CS:GO) in Katowice in 2018. In each match, there were 10 players, hence 10 player streams. There is 1 commentator stream from which participants are asked to find replays. The training data consist of recordings from 2nd March and 3rd March, while the test data are recordings from 4th March. Match metadata stores information about starting points and duration of matches in streams. There is also synchronisation data derived from game logs that show starting points of each round of matches [5]. Synchronisation data for commentator streams can be erroneous due to replays that create discontinuities in matches. The ground truth file shows location of replay segments in a commentator stream and source location of the replays in a player stream. The average duration of replays on the training data is 326 frames (roughly 5 seconds) and the longest replay segment is 843 frames (roughly 14 seconds).

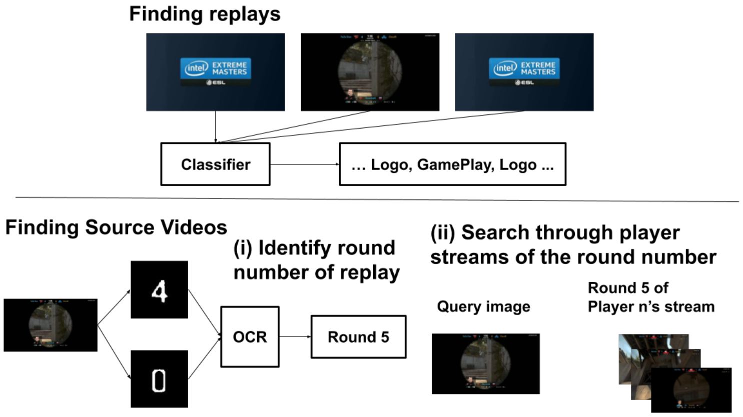
Finding replays
Classifier
... Logo, GamePlay, Logo ...
Finding Source Videos
4
0
(i) Identify round number of replay
OCR
Round 5
(ii) Search through player streams of the round number
Query image
Round 5 of Player n's stream

Figure 1: Two stage approach to finding replays and source segments of replays.

3 OUR APPROACH

We adopted a two stage approach to identifying replays in a commentator stream and finding their source location in a player stream.

3.1 Finding Replays

An algorithm for location of replays from a commentator stream needs to know a signal from which a replay begins. By analysing ground truth replays, we realised that these tend to begin with a sponsor logo. Therefore, our algorithm seeks to identify event patterns where a sponsor logo is followed by a game play segment.

To further analyse event patterns of replays, video frames were extracted from ground truth replay segments with 5 secs of preceding and following contexts at interval of 1 sec, which resulted in 1121 video frames. These video frames were transformed into fixed-length feature vectors using AlexNet [2] pre-trained on the ImageNet dataset [1], and k-means clustering was applied to the

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

vectors with the number of clusters set to 4. It was observed that replays and surrounding contexts can be classified into three categories: (i) game play, (ii) sponsor logo and (iii) audience or players.

To automatically classify video frames into these three categories, a neural network model with a single linear layer (equivalent to logistic regression) was trained on the feature vectors extracted from the video frames. Before training a model, the output of k-means clusters was manually corrected to ensure that video frames were not tagged incorrectly. The model was trained on 1093 video frames, with the remaining 28 video frames used for evaluating the model. The model showed 100% accuracy on the test set.

The trained neural network model was applied to segments of the commentator video corresponding to the matches according to the metadata. Since the model was trained on the video frames extracted from ground truth replays, when the model was applied to regions of a video outside matches, this is likely to cause mis-classification of video frames.

To obtain replay segments. First, segments starting with a sponsor logo, containing game plays, audience or players in the middle, and ending with a sponsor logo were gathered by scanning a sequence of output from the neural network model. Then, segments whose duration was less than 3 secs or longer than 20 secs were discarded, since the average duration of replays was roughly 5 sec. and the maximum was roughly 14 secs. Although some ground truth replays last for 7 or 29 frames, setting the minimum duration of replays to 0 secs led to worse results, most likely due to mis-classification of a short sequence of video frames.

3.2 Finding Source Videos of Replays

Our approach to finding source videos of replays relies on similarity of two video frames. However, it is prohibitively expensive to compute similarity of every replay with the whole player streams. Furthermore, if a battle field of a replay appears multiple times in multiple player streams, the system can be confused with wrong source videos. To reduce search space, first, a round number of each replay was identified by running OCR on score overlays, then the similarity of one video frame from replays with segments of a particular round number of matches in player streams was computed.

Since the synchronisation data of commentator streams can be erroneous due to replays. An alternative approach is to recognise match scores during replays (e.g, 1-4 means that the match is in the 6th round). To achieve this, video frames were extracted from the found replays at intervals of 1 sec. For each replay, regions of scores were cropped for an OCR system to recognise digits, then converted to a gray scale and binarised at a threshold of 110. For each replay, there were the same number of cropped images for each team. The final round number was decided based on a majority vote (i.e., when 5 cropped images were recognised as 1, 1, 1, 7, 7, this score was regarded as 1). When OCR failed to recognise digits from an image, the associated replay segment was discarded, since the replay was most likely a false positive.

Once a match number and a round number of every replay had been identified, one of the video frames was randomly chosen and compared to every 1 sec. of segments of player streams corresponding to the match number and the round number according to the synchronisation data. This step produced similarity scores of two video frames using mean squared error (MSE) and structural similarity (SSIM) [4]. The difference between these similarity metrics is that SSIM is more robust to noise in a target image. Video frames of player streams with the highest similarity score were considered as replay sources.

Table 1: Precision, recall and F1 scores of replay segments found by our approach.

	train			test		
	Precision	Recall	F1	Precision	Recall	F1
JC 0.5	0.82	0.58	0.68	0.69	0.50	0.58
JC 0.75	0.61	0.43	0.51	0.47	0.34	0.39

Table 2: The average overlap of source videos found by our approach with ground truth segments.

	train		test	
	MSE	SSIM	MSE	SSIM
JC 0.5	0.52	0.41	0.31	0.26
JC 0.75	0.56	0.45	0.33	0.18

4 RESULTS

Found replay segments were evaluated using precision, recall and F1 scores. The scores were calculated using the Jaccard index, where overlap of replay with ground truth segments is divided by union of replay with ground truth segments. When the Jaccard index is higher than the pre-defined threshold (0.5 and 0.75), prediction is considered as correct.

Table 1 shows results of the task to find replays from a commentator stream. On both training and test data, precision is higher than recall. As mentioned in Section 3.1, some of the ground truth replays are quite short and even though they begin with a sponsor logo, they do not end with a sponsor logo. It is likely that such replays were missed in our approach, accounting for lower recall.

Table 2 summarises the average overlap of found source videos with replays. Both on the training and test data, MSE led to a better result than SSIM. On the training data, 29 out of 81 ground truth segments were predicted to be from the wrong source videos, which provides an explanation of the low average overlap on the training data, and even lower on the test data.

5 CONCLUSIONS

This paper describes the DCU-ADAPT participation in GameStory at MediaEval 2019. We employed a machine learning approach to convert a commentator stream into a sequence of events audio identify replay segments. To find source videos of the replays, we limited search space using OCR and compared video frames from replays to player streams using similarity metrics including MSE and SSIM.

ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

REFERENCES

[1] Jia Deng, Wei Dong, Socher Richard, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. 1097–1105.

[3] Mathias Lux, Michael Riegler, Duc-Tien Dang-Nguyen, Johanna Pirker, Martin Potthast, and Pal Halvorsen. GameStory task at MediaEval 2019. In *Proceedings of MediaEval 2019*.

[4] Zhou Wang, Alan C Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[5] Michael Wutti. Automated Killstreak Extraction in CS:GO Tournaments. In *Proceedings of MediaEval 2018*.