음악 분석을 이용한 클라이언트 중심의 키프레임 생성 시스템

무즈타바 굴람, 김선대, 박은수, 김승환, 유재성, 류은석
가천대학교 컴퓨터공학과
{mujtaba, ele7004, dmsehf804, whitekomani, poopoo96}@gc.gachon.ac.kr,
esryu@gachon.ac.kr

Client-driven Animated Keyframe Generation System Using Music Analysis

Ghulam Mujtaba, Seondae Kim, Eunsoo Park, Seunghwan Kim, Jaesung Ryu and
Eun-Seok Ryu
Department of Computer Engineering,
Gachon University, Korea

Abstract

Animated images formats such as WebP are highly portable graphics formats that are being used everywhere on the Internet. Despite their small sizes and duration, WebP image previews the video without watching the entire content with minimum bandwidth. This paper proposed a novel method to generate personalized WebP images in the client side using its computation resources. The proposed system automatically extracts the WebP image from climax point using music analysis. Based on user interest, the system predicts the genre using Convolutional Neural Network (CNN). The proposed method can easily integrate with streaming platforms such as YouTube, Netflix, Hulu, and others.

1. Introduction

Over the last few years, animated WebP images have quickly risen in popularity in different platforms such as instant messaging, online journalism, social media, etc. Different from other media formats, WebP images are unique in that they are spontaneous (very short in duration), have a visual storytelling nature (no audio involved) and shared by online users [1]. Previous methods for extracting animated image formats from videos have neglected the user preferences, attention, and other [2]. Despite their increasing popularity and unique visual characteristics, there is a remarkable dearth of scholarly work on animated images in the personalized recommendation system community.

In an attempt to provide personalized animated images for recommended or selected videos, a novel client-drive system proposed to generate WebP explicitly takes a user's interests into account. Towards this goal, firstly recommended video is selected from video playlist to generate WebP. The system categorizes users related music files using deep learning music genre classification model. The model is trained on a convolutional neural network using GTZAN dataset [3]. The system determines the climax point and estimates pitch from the audio file using a deep CNN operating on the time-domain signal.

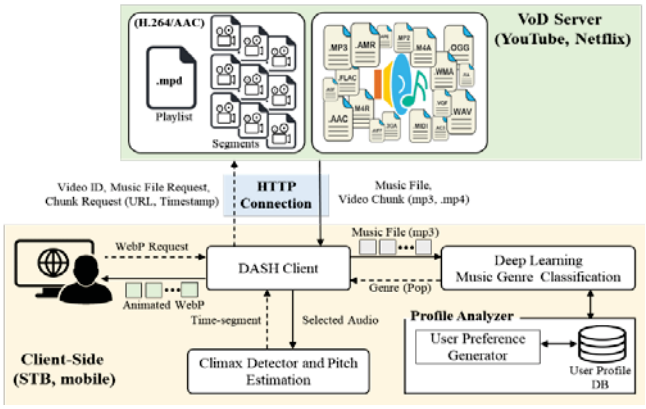

Figure 1: Conceptual system architecture

The corresponding frames are extracted from the video by estimating pitch position from the climax point to generate an animated WebP image.

We make the following contributions in this paper:

- An innovative client-driven method proposed to generate animated keyframe.
- Proposed a convolutional neural network to classify music genre on GTZAN Dataset.
- Estimated climax point from the music file and the estimated pitch.

The rest of the paper is organized as follows: Section II discusses the proposed approach for generation animated WebP. Section III includes discussion and concludes this work.

2. Proposed Approach

This section proposes the system approach for genre classification, climax detection, and pitch estimation. Figure 1 shows a conceptual diagram of the proposed system. This work only focusses on the client-side implementation. The main components of the client-side are deep learning music genre classification, climax detector and pitch estimator module, profile analyzer and DASH client.

2.1 Genre Classification:

The genre classification task conducted with GTZAN dataset [3], even though some drawbacks and limits are indicated [4]. GTZAN is still one of the widely used datasets as a benchmark for music genre classification [5]. The dataset consists of 1,000 song 30-second long music clips with the sampling rate of 22,050 HZ, 16 bits. Each clip is annotated with ten different genres and each genre there are 100 clips for the genres. The proposed system reads the audio files using melspectrograms [6] and split each clip of 30-seconds into 3 seconds windows with 50% overlapping resulting in a dataset with the size 19000x129x128x1 (samples x time x frequency x channels). The data is partitioned into two groups 70% for training and 30% testing. For genre classification, modified VGG16 convolutional neural network trained on Keras in backend on TensorFlow [7]. The designed network for music classification obtained 99% accuracy, and each experiment performed four times with approximately 2 hours training duration on GeForce RTX 208. Figure 2 presents the average results in the confusion matrix.

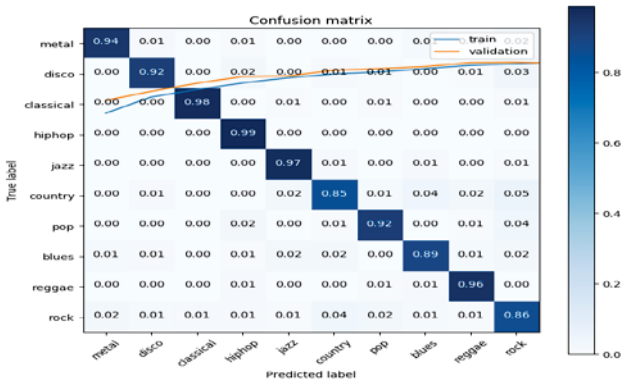

Figure 2: Confusion matrix using the mean

2.2 Climax Detection and Pitch Estimation

The plot structure of any movie is very important, it keeps users entertained until the video ends. According to Michael Hauge, there are five turning stages while planning the story structure of any movie, i.e. opportunity, change of plans, the point of no return, major setback and climax [8]. From these five phases, the most exciting and key part of any movie is the climax, where all the key events happen and turning the point of the story.

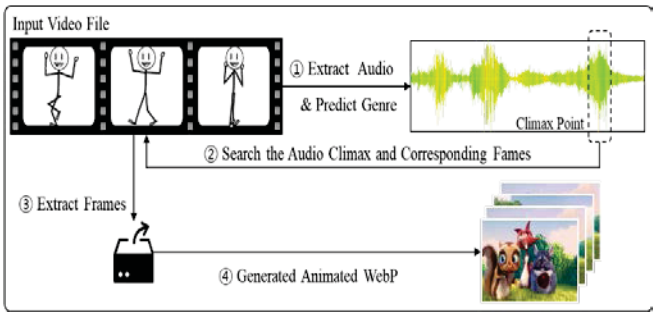

Figure 3: Overview of the system

Most movies follow the same structure and mostly climax part comes after 2/3 running time. Every climax has different music sound depends upon the genre of the movie such as action, disaster and war genres movies are loud at the climax part. On the contrary, the climax part in drama and romantic movies genres are quieter.

The proposed method gathers are user information in the client-side. According to user interests the climax point and pitch estimates from classification music file. The system extracts the climax points 75% length of the file to estimate pitch. The pitch estimates from climax point using state-of-the-art Convolutional Representation for Pitch Estimation (CREPE) [9]. CREPE uses a data-driven method for monophonic pitch tracking based on a deep convolutional neural network operating on the time-domain signal. With the highest pitch from the music file, that duration synchronizes to access specific video segment from DASH server. From a specific segment, the system uses FFmpeg to generate animated WebP image [10]. Generally, the length of the animated WebP in YouTube 3 seconds approximately. But the proposed method can extend the animated WebP. Figure 3 shows an overview of the system for generating animated WebP using music pitch estimation from the climax point.

3. Discussion and Conclusion

Animated WebP plays a very extensive role in streaming platforms such as YouTube, Dailymotion, Vimeo, etc. Users can easily preview a video even without watching the entire content. In previous researches for extracting animated images have neglect the user preferences, and music analysis. Users may easily miss their potentially favorite video while searching in the video playlist. Thus, WebP images play a very significant role to get users attention and more click relation for recommended videos.

This work proposes an innovative client-drive method to generate animated WebP images using music analysis. In addition, a deep learning genre classification model designed and trained in GTZAN dataset with 99% accuracy. The system automatically detects climax points from both video and audio files. The entire computations to process animated WebP images use client device resources. The server-side does not require extra computations to generate a WebP image. Thus, the proposed method easily integrates with streaming platforms such as YouTube, Netflix, Hulu, etc.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2017-0-01630)

References

[1] Bakhshi, Saeideh, et al. "Fast, cheap, and good: Why animated GIFs engage us." Proceedings of the 2016 chi conference on human factors in computing systems. ACM, 2016.

[2] Gygli, Michael, Yale Song, and Liangliang Cao. "Video2gif: Automatic generation of animated gifs from video." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[3] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing 10.5 (2002): 293-302.

[4] Sturm, Bob L. "The state of the art ten years after a state of the art: Future research in music information retrieval." Journal of New Music Research 43.2 (2014): 147-172.

[5] Sturm, Bob L. "An analysis of the GTZAN music genre dataset." Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. ACM, 2012.

[6] Librosa, Librosa GitHub page, "Python library for audio and music analysis" https://github.com/librosa/librosa. Accessed [May 2019].

[7] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[8] Press Release, Movie Outline, "The Five Key Turning Points of All Successful Movie Scripts", Movie Outline, URL: http://www.movieoutline.com/articles/the-five-key-turning-points-of-all-successful-movie-scripts.html, Last accessed [May 2019]

[9] Kim, Jong Wook, et al. "CREPE: A convolutional representation for pitch estimation." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[10] FFmpeg. FFmpeg GitHub page. https://github.com/FFmpeg/FFmpeg. Accessed [May 2019].