Automatic Video Intro and Outro Detection on Internet Television

Maryam Nematollahi
Dept. of Electrical and Computer Engineering
Ryerson University
Toronto, Ontario M5B 2K3
maryam.nematollahiar@ee.ryerson.ca

Xiao-Ping Zhang
Dept. of Electrical and Computer Engineering
Ryerson University
Toronto, Ontario M5B 2K3
xzhang@ee.ryerson.ca

ABSTRACT

Content Delivery Networks aim to deliver multimedia content to end-users with high reliability and speed. However, the transmission costs are very high due to large volume of video data. To cost-effectively deliver bandwidth-intensive video data, content providers have become interested in detection of redundant content that most probably are not of user's interest and then providing options for stopping their delivery. In this work, we target intro and outro (IO) segments of a video which are traditionally duplicated in all episodes of a TV show and most viewers fast-forward to skip them and only watch the main story. Using computationally-efficient features such as silence gaps, blank screen transitions and histogram of shot boundaries, we develop a framework that identifies intro and outro parts of a show. We test the proposed intro/outro detection methods on a large number of videos. Performance analysis shows that our algorithm successfully delineates intro and outro transitions, respectively, by a detection rate of 82% and 76% and an average error of less than 2.06 seconds.

Categories and Subject Descriptors

[**Media Transport and Delivery**]: Multimedia content-aware pre-fetching and caching, multimedia analysis and recommendations for media distribution and caching purposes; [**Multimodal Analysis and Description**]: Multimodal semantic concept detection, object recognition and segmentation

Keywords

Intro and Outro Detection; Media Streaming; Media Content Delivery

1. INTRODUCTION

Media streaming refers to content delivered over the Internet and requires a content delivery network (CDN) to distribute and deliver the content. A CDN is a large distributed system of servers deployed in multiple data centers across the Internet. The goal of a CDN is to serve content to end-users with high availability and high performance [7].

Multimedia content has a large volume. Despite advancements of compression techniques, media transmission costs are still significant. Content providers such as media companies (e.g. Netflix) have to pay CDN operators (e.g. Advection) for delivering their content to their audience of end-users. Enhancing online performance is critical for content providers because content delivery delays diminishes user's satisfaction of online watching experience and gradually reduces number of users. However, High speed delivery and data reliability incur high costs.

To seamlessly deliver bandwidth-intensive multimedia content and yet save costs, content providers have become interested in developing ways of detecting redundant content that are likely not to be of user's interest and stopping their delivery or delivering them on user's request and preferences, e.g. commercial detection and removal methodologies [5]. In this work, we target intro and outro segments of a video which, by definition, are the opening and closing sections of a TV program, respectively. Intros are made up of the title shot, opening credit, music and shots meant to artistically give a glimpse of main theme of the show to set the audience's mood for it. Outros at the end of programs are made up of closing credits and music to list the production cast.

Customarily, all episodes of a typical TV series have exactly the same intro and outro. Once made, intro and outro parts are duplicated and inserted into all the episodes of a TV show. No matter how creative and memorable the intro and outro are, viewers get bored with what is repeated every time they watch that show; they find it a tiresome waste of their time to be presented with 30 seconds to 3 minutes of repetitive content; they usually fast-forward to skip it and watch the main story. To moderate the annoyance of repetitive content in mainstream TV, the outro time duration was often used to promote other shows on the network.

Considering high costs of delivering multimedia data in Internet media streaming, it would be reasonable to avoid delivering the data that is uninteresting and annoyingly time-consuming; e.g. repetitious intro and outro segments of video. That is, it is intelligent to save costs over unwanted content. Consider the animated series "Tabaluga" as an example which has 80 episodes. Each episode of Tabaluga is about 24 minutes long. It has an intro length of 90 seconds and an outro length of 90 seconds as well. Therefore, where the frame rate of the video is 25 frames per seconds, frame width is 656 pixels and frame height is 480 pixels, the whole Tabaluga series contain $80[series] \times 180[seconds] \times$

WISMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2661714.2661729.

62967 kbits of redundant data regarding the intro/outro (IO) segments. Thus, we propose to develop an IO detection methodology that automatically identifies and skips these undesirable segments of video.

Recently, video game producers have provided the "Skip Intro" option, e.g. for video games such as "Batman Arkham Origins" and "Borderlands". However, to the best of our knowledge, there has not been any methodology proposed for IO identification regarding Internet media streaming. Although commercial detection and near duplicate detection problems resemble the IO identification problem in nature, the methodologies applied to them in literature cannot be applied to IO identification problem. Successful commercial detection algorithms require the presence of a series of broadcast editing rules such as absence of TV station logos or subtitles [1, 4]. Other commercial detection studies are mostly based on distinctive features of commercials in comparison with main video stream in terms of audio-visual characteristics such as music and speech color difference, and also drastic illumination and scenery changes [5, 10]. IO identification is also different from near-duplicate detection problem [9, 13] because in the latter, we look for identicals of a known segment of video from among a large collection of videos. Regarding IO identification, we aim to detect this segment without access to reference video.

In this paper, to identify the location of transition from intro to main video content and location of transition from main video to outro, we rely on intrinsic characteristics of intro and outro such as silence gap, blank dark screen transitions and their repetitive nature over all episodes of a show. We apply signal processing methodologies to formulate this characteristics and identify intro and outro. We perform audio-visual feature extraction, frame-level histogram calculation, shot boundary detection [2], filtering and thresholding to extract intro and outro segments. We develop the IO detection algorithm for both single-episode and multiple-episode shows. We test the proposed framework on a large number of video series. Performance analysis shows that our algorithm successfully identifies intro and outro transitions with a detection rate of 82% and 76%, respectively; and the average offset error of our methods is less than 2.06 seconds.

2. INTRO DETECTION

2.1 Single Episode Show

We rely on blank dark screen transition and silence gap transition properties at the intro and outro location to detect them using only one single episode. In the following, we define each of these properties and describe how to calculate them. We look for these properties in the first (last) 3 minutes of the video to indicate intro/outro transition location.

2.1.1 *2.1.1- Dark Screen Transition:*

A common characteristic of location of transition from intro/outro to main video is that the user experiences a short sequence of blank dark frames. This phenomenon is clearly shown in a frame level in figure 1 from "Altair in Starland" cartoon. To detect blank screen transition property, we perform frame-level feature extraction. We apply Color histograms to represent the distribution of colors in a frame, that is, the number of pixels in the frame regarding each possible color. Color histograms of two similar frames are assumed to have similar counts of most colors. However,

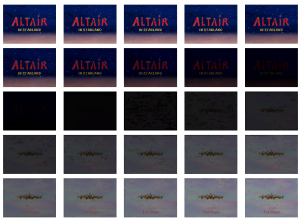

Figure 1: Dark Screen Transition Example.

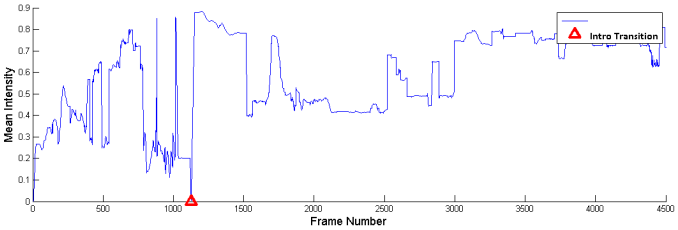

Figure 2: Mean Intensity of all frames of the first 3 minutes of one episode of Altair.

object motion and camera motion disturb this assumption to some extent so that they do not necessarily match on a per pixel basis [3]. We compute color histograms of all consecutive frames throughout the video to detect the shot boundaries. Then, for localizing blank screen shot transition occurences, we calculate the mean gray-scale intensity value $\overline{I_k}$ of each video frame (k):

$$\overline{I_k} = \frac{1}{N}\left(I_1 + I_2 + ... + I_N\right) \quad (1)$$

where I_i is the gray intensity value of pixel i, N is the total number of pixels. Figure 2 shows the sequential result of mean intensity value calculation on all frames of first 3 minutes of one episode of "Altair in Starland" cartoon. Intro duration of this cartoon is 45 seconds. As illustrated, a minimum happens to mean intensity at location of transition from intro to the main video which is around the frame number 1125 since the frame rate is 25 frames per seconds.

2.1.2 *2.1.2- Silence Gap:*

Another property of IO transition time is that the video falls into utter silence. Therefore, we need to study the audio signal carried over the video. An example of the silence gap at location of transition from intro to main video is given in figure 3 by illustrating the amplitude of the audio signal which is sampled at a rate 48000 Hz [3]. For silence detection, we need to consider the audio signal only in the time domain. To approximate the loudness or volume of a sound, we apply root mean square (RMS) of the signal energy [8]. We also use zero crossing rate (ZCR) which is defined as the number of changes in signal amplitude sign in the current frame; the higher the sound frequency, the higher is the ZCR value. It has been proved that music has a lower variability of the ZCR than speech [3]. IO parts of the video generally contain constant high-volume music, whereas the audio signal over the main video is composed of mainly speech and occasional low-volume music. To detect the silence transitions, we apply the loudness and ZCR so that if they are both below a certain threshold, we conclude that this frame may be silent. We only consider the first (last) detected silence gap in time to be the intro (outro) candidate if it lasts for at least half a second.

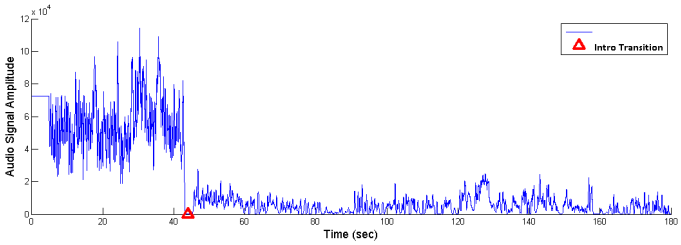

Figure 3: Silence gap at intro time (Altair).

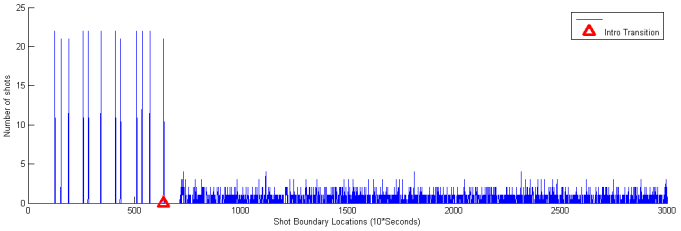

Figure 4: Shot Boundary alignment over 25 episodes of "Maja the Bee" series.

2.2 Multiple Episode show

Regarding multiple episode shows, we make use of the repetition of the intro (outro) in the beginning (ending) of all episodes of a TV series to detect them. For this purpose, we rely on detecting shot boundaries in the video sequence. Since the intro clip is the same at the beginning of all episodes of a series, the same shot boundary occurrence timing is repeated during intro time in all of them. However, as the video moves to the main story, everything is new; therefore, shot boundaries occurrences serve the progression of the new story and does not follow any pattern.

To detect the shot boundary transition locations throughout each video, we apply frame-level histogram features [2, 11, 6]. We assume that color histograms differences of consecutive frames within the same shot are smaller than the difference between histograms of frames belonging to different shots [11]. Having computed the histogram of all frames, we apply the independent component analysis (ICA) method as proposed in [12] to delineate the shot boundaries since their approach has been shown to effectively detect both abrupt transitions and gradual transitions of shots.

We look for the intro (outro) transition time in the first (last) 3 minutes of a show by studying the histogram of the shot boundary transition timings. We consider a time resolution of $\frac{1}{3[minutes]\times 60[seconds]\times 10}$ for possible times of shot boundary occurrence from 0 to 3 minutes. That is, we empirically consider 10 possible bins in each second. Furthermore, we define 1800 number of bins corresponding to these possible times and initialize them with count zero.

Considering M episodes of a show and all the detected shots throughout these videos, we build a histogram which counts the number of shot boundary observations that fall into each of the disjoint 1800 time instants (bins). For each shot, we count its corresponding time bin up by one. If we let n to be the total number of shots throughout this M 3-minute videos, then $n = \sum_{i=1}^{k} n_i$. where k is the total number of bins and n_i is the number of shots in i^{th} bin. Having built the histogram of shot frequencies for the 1800 time instances, as it is shown in figure 4, it is evident that the repetitive pattern of intro shots causes formation of large peaks at their corresponding occurrence times. The intro transition time is when the last peak happens (figure 4) since as we go through the main video, shot boundaries are planned based on a new story and do not repeat in following episodes.

To delineate the intro transition time, we define a threshold value (θ_{intro}) which we can use to distinguish shot-frequency peaks in the histogram. Since the peaks are formed by counting the number of times that the same shot happens regarding the M episodes, it is reasonable to set this threshold to M. However, sometimes producers add a few blank frames to beginning of a video to shift it for a fraction of a second for either synchronization or encoding purposes. Hence, the intro shots of different episodes might not be perfectly synchronized so that the occurrence time of the same shot may fall into not same bin but a close by bin. To compensate for this alignment inaccuracy, we perform moving average operation on the histogram as follows to create a sequence of averages of different subsets of the full histogram:

$$n_i^{'} = \frac{1}{l}\left(n_{i-\frac{l}{2}} + n_{i-\frac{l}{2}+1} + ... + n_i + ... + n_{i+\frac{l}{2}-1} + n_{i+\frac{l}{2}}\right)$$

where $n_i^{'}$ is the number of shots in bin i after being averaged. We fixed the averaging window length to $l = 10$ bins. In above equation, centered averaging ensures mean variations are aligned with the variations in the data and not shifted in time. Due to the moving average operation, one might need to adjust the threshold value for discriminating shot-frequency peaks depending on the averaging window length (l). In this work, since we chose a small value for the averaging window length, selection of threshold value $\theta_{intro} = M$ worked well to delineate the location of last intro shot which accordingly is the time when the video makes transition from intro part to the main story.

3. OUTRO DETECTION

Outro detection methodology is very similar to intro detection. For delineating the outro transition time from main video to outro, we consider the last 3 minutes of M episodes of a TV series (outro segments are often shorter than intros). For outro detection based on the multiple episode method, we adjusted the threshold value, θ_{outro}, for discriminating shot-frequency peaks in outro histogram to $max(M, mean(q), std(q))$; where q is the set of all non-zero values in the histogram and *mean* and *std* stand for average and standard deviation of q. This modification was necessary due to some minor asynchronicity in video endings that cause inaccurate alignment of shots of different episodes.

4. PERFORMANCE ANALYSIS

To evaluate the proposed intro/outro detection framework, we run two separate set of experiments regarding single-episode and multiple-episode methods. We have a total of 27 animated TV series in our database whose length ranges from 5 to 25 minutes. The database contains old as well as recently-produced series. Video intros vary from 13 to 135 seconds; outro duration variation range is from 10 to 123 seconds. Video series names and their corresponding intro/outro duration are listed in table 1. For the last 6 series, experiments were only done for single episode case since we downloaded one episode of them from Youtube and did not have access to full series. In table 1, true values of intro and outro duration, T_i and T_o, are given on 3^{rd} and

Table 1: Intro and Outro detection results on a collection of children video series (x: miss-detection).

No.	Video Name	intro true (T_i) (Secs)	Intro result (D_o) Single Episode	Intro result (D_o) Multiple Episodes	Outro true (T_o) (Secs)	Outro result (D_o) Single Episode	Outro result (D_o) Multiple Episodes
1	ABC Monsters	47	45	48	57	8 **x**	57
2	Animal Atlas	50	48	50	80	78	77
3	Safari Tracks	41	43	42	31	8 **x**	29
4	Willy Fog	104	103	102	95	123 **x**	92
5	David The Genome	73	73	73	90	85	90
6	The Mozart Band	109	80 **x**	108	45	46	115 **x**
7	The Brothers Flub	67	69	66	60	61	57
8	Raindrop	93	52 **x**	92	47	45	46
9	Florries Dragon	70	45 **x**	68	25	25	18 **x**
10	Kerwhizz	45	55 **x**	54 **x**	32	6 **x**	22 **x**
11	Lapitch	90	90	88	90	90	89
12	Maja The Bee	71	71	72	58	82 **x**	57
13	Tabaluga	90	92	65 **x**	90	90	86
14	Vipo	54	35 **x**	54	23	7 **x**	21
15	Wilf the Witchdog	45	45	45	15	15	120 **x**
16	Altair in Starland	45	44	45	15	15	14
17	Fluffy Gardens	21	20	32 **x**	20	20	16
18	Larry Lawnmower	35	36	35	25	5 **x**	22
19	Insect Antics	22	23	20	10	19 **x**	7
20	Meg and Mog	13	13	13	16	17	9 **x**
21	Urmel	35	36	64 **x**	34	43 **x**	32
22	Vanguard	135	123 **x**	—	123	110 **x**	—
23	Turtle Hero	85	85	—	60	60	—
24	Spider Man	65	60	—	30	30	—
25	Bugs Adventure	80	81	—	65	63	—
26	Black Night	40	35	—	25	20	—
27	Innocents Abroad	80	77	—	37	38	—

Table 2: Overal IO detection performance results

Task	Detection Rate Single Episode (%)	Detection Offset (seconds)	Detection Rate Multiple Episode (%)	Detection Offset (seconds)
Intro Detection	**77**	**1.42**	**82**	**0.82**
Outro Detection	**63**	**1.17**	**76**	**2.06**

6^{th} columns from left. Regarding intro detection results, column 4 and 5 present the results (D_i and D_o) of single- and multiple-episode detection methods; respective Columns 8 and 9 report outro detection results.

If detection offset is less than 5 seconds, $|D_{i/o} - T_{i/o}| < 5$, then we consider a detection result $D_{i/o}$ to be correct. Overal detection statistics from table 1 are presented in table 2. Applying single-episode method, intro transition time is correctly delineated for 77% of total number of videos with average 1.42 seconds of offset inaccuracy. Using multiple-episode method, intro detection rate increased to as high as 82% and the offset value was improved to 0.82 seconds. Regarding outro, experiments show that multiple-episode method is more accurate than the single-episode method by 13% (table 2). A detected intro/outro value that differs the true value by more than 5 seconds is considered to be a missed detection (x). For some of the missed detections of single-episode method, there actually has not been a silence gap or dark screen transition to be identified, e.g. outro of Kerwhizz series. For these kinds of videos, multiple-episode method is the more appropriate detection method.

Recently, producers are becoming interested in making intro clips whose design are creatively updated to either coordinate with the main story theme for each episode. Delineation of intro for these videos will be challenging using the multiple method. In future work, we intend to circumvent this problem using training-based classification methods.

5. CONCLUSION

Due to growing popularity of Internet Television and yet-expensive costs of streaming bandwidth-intensive large-volume video data, it is reasonable to avoid delivering redundant content that most probably are not of user's interest. In this paper, we propose two computationally-efficient methods to automatically segment intro (opening credit) and outro (ending credit) which are tediously repeated in all episodes of TV shows to provide an option to purposefully skip them. For this purpose, we rely mainly on as straightforward features of the video as frame-level histogram, shot boundaries, silence gaps and blank screen transitions. For validation studies, we perform our experiments on a large collection of video series. Performance analysis show that the proposed method identifies intro and outro segments successfully by a recognition rate of 82% and 76%, respectively, with an average offset error of less than 2.06 seconds.

6. REFERENCES

[1] A. Albiol, M. J. Ch, F. A. Albiol, and L. Torres. Detection of tv commercials. In *IEEE Intl. ICASSP Conference Proceedings*, pages 541–544, May 2004.

[2] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5(2):122–128, April 1996.

[3] D. Brezeale and D. J. Cook. Automatic video classification: A survey of the literature. *IEEE Trans. on Systems and Cypernetics - PART C: Applications and reviews*, 38(3), May 2008.

[4] Y. P. Huang, L. W. Hsu, and F. E. Sandnes. An intelligent subtitle detection model for locating television commercials. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(2):485–492, Apr. 2007.

[5] N. Liu, Y. Zhao, Z. Zhu, and H. Lu. Exploiting visual-audio-textual characteristics for automatic tv commercial block detection and segmentation. *IEEE Trans. on Multimedia*, 13(5):961–973, 2011.

[6] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *8th ACM Intl. Conf. Multimedia*, pages 219–227, September 2000.

[7] A. Vakali and G. Pallis. Content delivery networks: Status and trends. *IEEE Internet Computing*, 7(6):68–74, 2003.

[8] E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, 3(3):27–36, Fall 1996.

[9] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan. Real-time near-duplicate elimination for web video search with content and context. *IEEE Trans. on Multimedia*, 11(2):196–207, 2009.

[10] S.-H. Yang, C.-W. Fan, and Y.-C. Chen. An improved automatic commercial detection system. *Visual Communications and Image Processing*, pages 1–4, 2011.

[11] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1:10–28, 1993.

[12] J. Zhou and X. P. Zhang. Video shot boundary detection using independent component analysis. In *IEEE Intl. ASSP Conference Proceedings*, March 2005.

[13] X. Zhou, X. Zhou, L. Chen, A. Bouguettaya, N. Xiao, and J. A. Taylor. An efficient near-duplicate video shot detection method using shot-based interest points. *IEEE Trans. on Multimedia*, 11(5):879–891, 2009.