NII_UIT at TRECVID 2022: Movie Summarization Task

Nam Nguyen[1], Tien Hung Nguyen[1], Cong Nguyen Thanh[1], Hao Vo[1], Khiem Le[1]
Tien-Dung Mai[1], TruongAn PhamNguyen[1], Duy-Dinh Le[1], Shin'ichi Satoh[2]

[1] University of Information Technology, VNU-HCM, VietNam

[2] National Institute of Informatics, Japan

November 16, 2022

Abstract

Collecting key facts about a movie character in a full length movie is the all new and shinny track in TRECVID 2022. We built upon our approach for TRECVID 2021 Video Summarization Task, that fused together both visual and auditory cue in the original movie to face this challenge. Though the results leave a lot to be desired, our early findings may served as good starting points for others.

1 Introduction

This report details our participation in TRECVID 2022 Movie summarization track [ACB+22]. This track replaces the Video Summarization Task of prior years and aims to capture important facts about certain persons during their role in the movie storyline. The input of this tasks is a full length movie, a character in that movie along with image or video examples of that character. We are to collect important and critical events about that character storyline. Those key-facts are expressed through either a video summary with limited maximum length or a textual summary with limited number of words and sentences.

The first challenge for this tasks is that visual information is not enough to generate good summary. For a movie character may appear in wildly different scene with wildly different makeup. There are many cases that even state-of-the-art facial recognition system failed to recognize target character in the video sequence. Moreover, key-fact about a character may be disclosed when that character is not visible on the screen. A solution using visual cue as the sole driving force would, therefore, be doomed to miss important fact. The second challenge is the difficulty in processing audio information from movie. The background music and sound added for drama effect hampered audio-to-text system. Available online transcript may provide clear text of dialogues but lacked information about who was saying that a particular line in those dialogues. And finally, even if you can get a hold of the original scripts of the movie, it's still extremely difficult for computer to understand the storyline through dialogue and determine whether a piece of text contain any key-facts.

2 Our Approach

The pipeline of our approach is given in 1 Each movie shot will be passed through the pipe to calculate the 'face similarity ranking score' and 'text similarity ranking score'. Those scores would then be fused together to generate a single important score for the shot, which will determined whether or not that shot is included in the final summary.

2.1 Shot Splitting

Compare to last year Video Summarization Task, this year video dataset does not provide pre-defined segment that need to be select into summary. On the contrary, participants can select any arbitrary

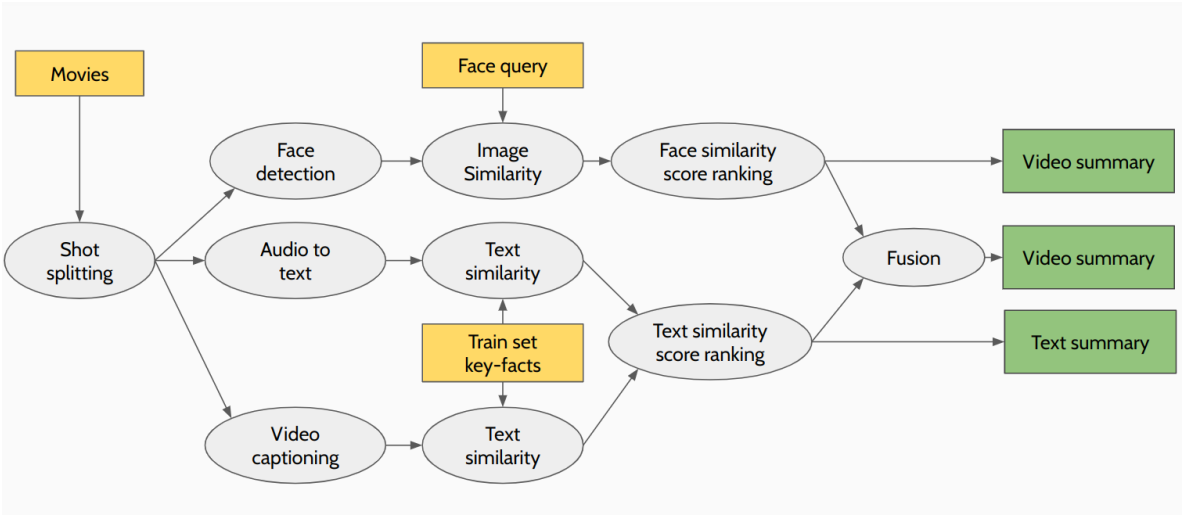

Figure 1: Our overall pipeline

segments of the original movie into the final summary. In order to build upon the foundations from our past years' approach, we first have to split the movie into a collection of short shots.

We use TransNet_v2 for this video segmentation task. Since a majority of video segmenting systems was built to segment video based on visual effect, not storytelling we had to filter too long segment that would not fit into the constraint of output summary.

2.2 Face Similarity Ranking Score

We reused the baseline from our TRECVID 2021 participation [KDTS21] to calculate this score. MTCNN [ZZLQ16] is used for face detection, VGGFace2 [2] is sued for face representation. Finally cosine similarity is to match between the faces in the shot and the input face query, generating the face similarity ranking score:

$$sim(query, shot_i) = \frac{1}{N}\sum_{k=1}^{N}(max_{j=1,2,\cdots,M}(cos(desc_k^{query}, desc_j^{shot_i}))$$

where N is the number of faces of character in the input and M is the number of faces in the current shot. The notion $desc_k^{query}$ means the descriptor vector of the k-th face in the query and the notion $desc_k^{shot_i}$ means the descriptor vector for j-th face in the i-th shot.

2.3 Text Similarity Raking Scores

There are two sources for our text similarity score.

2.3.1 Audio to Text Similarity Score

First we passed each shot through Video captioning network. The we compare the generated caption with the textual key-facts in training sets. Using Bert as feature extractor and cosine distance as text similarity measurement. This generate $audio_{score}$

2.3.2 Video Captioning Similarity Score

The second sources for text similarity come from the audio in the shot. We passed each shots through Audio to text system. The generated audio caption went through the same process as the video caption. This two texts will be the key ingredient to generate textual summary as well as the fusion formula for our overall shot important score. This generate $caption_{score}$

Table 1: Training set

video_name.character	time_of_movie	# scene	time_of_scene (s) min	time_of_scene (s) max	time_of_scene (s) avg	# key-facts
Calloused_Hands.Byrd	**1:37:16**	**65**	20	**247**	**96**	**35**
Calloused_Hands.Debbie						24
Liberty_Kid.DERRICK	1:31:42	56	**12**	299	**94**	27
Llike_me.Burt_Walden	1:23:56	**28**	**47**	**300**	**167**	**4**
Like_me.Kiya						12
losing_ground.Sarah_Rogers	1:25:38	**40**	29	246	**120**	15
Memphis.willis	**1:18:39**	47	17	294	97	13
Mean						26

Table 2: Test set

video_name	time_of_movie	# scene	time_of_scene (s) min	time_of_scene (s) max	time_of_scene (s) avg	Duration (s)
Archipelago.Cynthia	**1:50:04**	**57**	21	**389**	**113**	190
Archipelago.Edward						140
Bonneville.Arvilla	1:32:39	41	19	269	124	190
ChainedforLife.Mabel	1:29:28	38	**15**	370	136	130
heart_machine.Cody	1:23:37	**28**	22	451	**158**	160
heart_machine.Virginia						110
Little_Rock.Atsuko	**1:22:48**	39	24	289	121	190
Little_Rock.Cory						160
Mean	1:31:43	40.6	20.2	353.6	130.4	138.75

2.4 Shot Time and Tempo

We devised a time score with the formula $time_{score} = 1 - \frac{time_{shot}}{time_{movie}}$ to penalized too long shot. It was our experience that a summary made up of many short shots usually out perform summary from long shots.

2.5 Fusion Score

To simplify the selection process to determine which shot would end up in the summary, we combined above score into a fusion important score

$$important_{socre} = \begin{bmatrix} face_{Score} \\ audio_{score} \\ caption_{score} \\ 1 - \frac{time_{shot}}{time_{movie}} \end{bmatrix} * \begin{bmatrix} w_{face} & w_{audio} & w_{caption} & w_{time} \end{bmatrix}$$

Where $W = \begin{bmatrix} w_{face} & w_{audio} & w_{caption} & w_{time} \end{bmatrix}$ is the fusion weights vector, with the constraint that $w_{face}+w_{audio}+w_{caption}+w_{time} = 1$ We use these weight to estimate the impact of each individual score to the performance of final summary.

The next section will discuss at length our experiment process during the four runs with TRECVID assessors.

3 Datasets and Experiments

The datasets provided consists of one training set and one test set of 5 movie each. Some statistics about the data set is given in table 1 and table 2

At first glance one can notice the extremely strict constraint on output summary duration from the test sets. A main character storyline may consists of more than 30 keyfacts (inferred from training sets) and conveys all that information in a summary video of just a little more than 2 minutes is extremely challenging even for professional human.

We came into the task knowing full well that good performance may not be possible, so our 4 runs was used for water testing purpose, to attest the impact of different apsects of our approach.

First run was use with fusion score coming entirely from $face_{score}$, i.e $W = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}$, 2nd and 3rd run have fusion weights that prioritized text score, and our last run was a bit more balanced weights but $face_{score}$ was still more prominent.

4 Result and Discussion

Even though our results was far from perfect, we got some insights about this challenging task. Despite the obvious fact pointed out by task organizer that key-fact event may be disclose when target character is not visible on screen. Our experiment show that $face_{score}$ still contribute mo prominently to the final performance.

We suspect the cause of that is the main device for story telling in movie is through dialogue between characters that sometimes even involved facial expression and body language. This level of semantic understanding is extremely hard for computer, even when having full transcript of the movie available. Thus, despite of our best effort to extract meaning from the text of the movie, both $caption_{score}$ and $audio_{score}$ fail to represents the required key-facts.

References

[ACB+22] George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, , and Georges Quénot. An overview on the evaluated video retrieval tasks at trecvid 2022. In *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[KDTS21] Tien Do Tien-Dung Mai An Pham Nguyen Truong Duy-Dinh Le Khang Dinh Tran, Nhat Pham Le Quang and Shin'ichi Satoh. Nii-uit at trecvid 2021: Video summarization task. In *2021 TREC Video Retrieval Evaluation*, 2021.

[ZZLQ16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.