Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization

ISMAIL HARRANDO*, ALISON REBOUD*, PASQUALE LISENA, and RAPHAËL TRONCY, EURECOM, France
JORMA LAAKSONEN, ANJA VIRKKUNEN, and MIKKO KURIMO, Aalto University, Finland

This paper describes a fan-driven and character-centered approach proposed by the MeMAD team for the 2020 TRECVID [Awad et al. 2020] Video Summarization Task. Our approach relies on fan-made content and, more precisely, on the BBC EastEnders episode synopses from its Fandom Wiki[1]. This additional data source is used together with the provided videos, scripts and master shot boundaries. We also use BBC EastEnders characters' images crawled from the Google search engine in order to train a face recognition system. All our runs use the same method, but with varying constraints regarding the number of shots and the maximum duration of the summary. The shots included in the summaries are the ones whose transcripts and visual content have the highest similarity with sentences from the synopsis. The runs submitted are as follows:

- MeMAD1: 5 shots with highest similarity scores and the total duration of the summary is < 150 sec;
- MeMAD2: 10 shots with highest similarity scores and the total duration of the summary is < 300 sec;
- MeMAD3: 15 shots with highest similarity scores and the total duration of the summary is < 450 sec;
- MeMAD4: 20 shots with highest similarity scores and the total duration of the summary is < 600 sec.

Surprisingly, the scores obtained for each run are very similar for the questions answering part of the evaluation. One exception concerns the character Ryan, for which one additional question is answered when choosing at least 15 shots. For all the runs, the redundancy score improved with the number of shots included in the summary while the relation with the scores for tempo and contextuality seem to vary more. The scores are lower for the question answering evaluation part. This is rather unsurprising to us as we realized while deciding on a similarity measure score that it is challenging for humans to choose between two potentially interesting moments without knowing beforehand the questions included in the evaluation set. Overall, we consider that the results obtained speak in favour of using fan-made content as a starting point for such a task. As we did not try to optimize for tempo and contextuality, we believe there is some margin for improvement. However, the task of answering unknown questions remains an open challenge.

ACM Reference Format:
Ismail Harrando, Alison Reboud, Pasquale Lisena, Raphaël Troncy, Jorma Laaksonen, Anja Virkkunen, and Mikko Kurimo. 2021. Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization. In *Proceedings of TRECVID 2020, International Workshop on Video Retrieval Evaluation (TRECVID 2020).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.
[1]https://eastenders.fandom.com/wiki/EastEnders_Wiki

TRECVID 2020, December 8–11, 2020, Virtual Conference

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Considering video summarization as an important task for digital content retrieval and reuse, the TRECVID [Awad et al. 2020] Video Summarization Task (VSUM) 2020 aims at fostering the research in the field by asking its participants to automatically summarize "the major life events of specific characters over a number of weeks of programming on the BBC EastEnders TV series"[2]. More precisely, for three different characters of the series, the participants have to submit 4 summaries with respectively 5, 10, 15 and 20 automatically selected shots. These generated summaries are evaluated by the assessors according to their tempo, contextuality and redundancy as well as with regards to how well they contain answers to a set of questions unknown to the participants before submission. In addition to the videos, the episodes transcripts are provided by the organizers.

We propose a character centered content summary approach based on fan-written synopses. The approach relies on scraping the Fandom EastEnders Wiki content for the episode synopsis and casting, in order to align them with the corresponding episodes. We include the shots that obtain the best similarity score with a sentence from the synopsis in our runs.

2 APPROACH

Our fan-driven and character centered approach is presented in Figure 1.

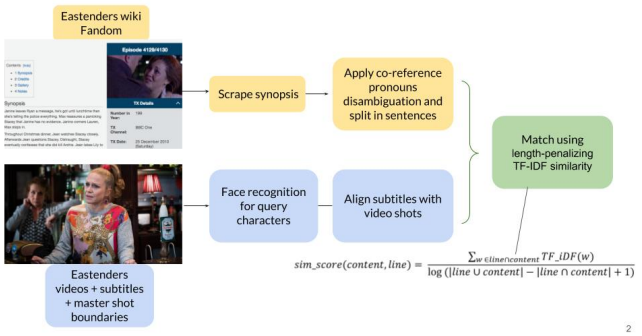

Fig. 1. Fan-driven and character centered approach

[2]https://www-nlpir.nist.gov/projects/tv2020/vsum.html

2.1 Scraping Synopses From the Fandom Wiki and Selecting Shots

The first step of our approach consists in scraping synopses available on the Fandom EastEnders Wiki[3].

Our main hypothesis is that every sentence (ending with a period) represents an important event to be added to the final video summary. We scrape the Synopsis and the Cast sections for each episode broadcasted between the dates of the provided episodes. The mapping between the episodes and their dates is in `eastenders.collection.xml` provided by the challenge organizers.

In parallel, we extract the shots in which the three characters of interest appear from the video. We run the Face Celebrity Recognition library[4], a system that relies on pictures crawled from search engines using the actor's name as search keyword. In our experiments, we have added "EastEnders" to the character names in order to avoid retrieving pictures of different people with the same name. For each picture, faces are detected using the MTCNN algorithm and the FaceNet model is applied to obtain face embeddings. Following the assumption that the majority of faces are actually representing the searched actor, other faces – e.g. person portrayed together with the actor – are automatically filtered out by removing outliers until the cosine similarity of face embeddings has a standard deviation below a threshold of 0.24 which has been empirically defined.

The remaining faces are used to train a multi-class SVM classifier, which is used to label the faces detected on frames. For more consistent results between frames, the Simple Online and Realtime Tracking algorithm (SORT) has been included, returning groups of detection of the same person in consecutive frames.

We select the shots displaying any of the the three characters of interests, keeping only those detection having a confidence score greater than 0.5. We also tried to use speaker diarisation to corroborate the visual information about the characters. However, given the limitations of the current technologies in terms of number of characters and the difficulty of identifying the character corresponding to each voice, we could not pursue the idea further.

2.2 Synopses and Transcript Pre-Processing

A synopsis for each episode was created using the provided files *eastenders.collection.xml* and *eastenders.episodeDescriptions.xml*. Since these were "EastEnders Omnibus" episodes, they correspond to multiple actual weekday episodes. We use the dates and the continuation to generate one synopsis for each "long" episode (typically made of 4 episodes). We then split the synopses into sentences and performed coreference resolution on the synopses to explicit character mentions using https://github.com/huggingface/neuralcoref. In parallel, the provided XML transcripts were also converted into timestamped text and aligned with the given shot segmentation. Finally, both the synopses sentences and shot transcripts were lower cased, stop words removed and lemmatized.

We also produced automatically-generated visual captions following the method presented by the PicSOM Group of Aalto University's submissions for the TRECVID2018 VTT task [Sjöberg et al. 2018]. The hypothesis is that by describing the visual information of a shot, visual captions could complement well the dialog transcript and therefore allow for a better matching between the shots and synopses sentences.

[3]https://eastenders.fandom.com/wiki/EastEndersWiki
[4]https://github.com/D2KLab/Face-Celebrity-Recognition

Table 1. Average score for each run and team

TeamRun	Percentage
MeMAD1	31%
MeMAD2	31%
MeMAD3	**35**%
MeMAD4	32%
NIIUIT1	9%
NIIUIT2	8%
NIIUIT3	8%
NIIUIT4	6%

2.3 Matching and Runs Generation

We perform a synopsis sentence / shot transcript pairwise comparison by generating a similarity score. We define similarity between two sentences as the sum of TF-IDF weights (computed on the transcript) for each word appearing in both of them, divided by the log length of the concatenation of both sentences, thus penalizing long sentences that match with many transcript lines.

Next, we order the shot by similarity score, picking only the best match for each shot (but not the other way around). This gives us scenes we are sure to appear in the summary, but not necessarily any guarantee about how important these scenes are. We also performed the pairwise comparison adding the automatically generated captions. A qualitative assessment revealed, however, that the captions were too noisy to complement well the transcript. We also make sure that if a line of dialog runs through the next shot, we include the next shot as well to improve the smoothness of the viewing. However, this heuristics was only relevant for the longest run (20 shots). Each run is made by selecting the N most matching shots out of the top, in chronological order.

3 RESULTS AND ANALYSIS

The final results for the two teams which have participated in TRECVID VSUM are presented in Table 1 while the detailed scores of our approach are presented in Table 2. Our method obtains the best overall score for each of the 4 required runs. The mean scores (range 1 - 7. High is best) for tempo, contextuality and redundancy are all above average (respectively 4.75, 4.75, 4.1) despite the fact that our method does not specifically attempt to optimise these metrics. However, in terms of question answering, the results show that the shots selected did not allow to answer more than two (at best) of the five questions. More specifically, Table 3 shows (in bold) the questions that were answered in at least one of ours runs. We notice that most of the questions started either with 'What' or 'Who' and that our approach performed equally for both types of questions.

4 DISCUSSION AND OUTLOOK

This paper describes a character centered video summarization method based on fan-made content, subtitles and face recognition.

Table 2. Detailed score for MeMAD's approach

Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5
Janine1	6	4	5	No	No	No	No	Yes
Janine2	5	5	6	No	No	No	No	Yes
Janine3	5	5	6	No	No	No	No	Yes
Janine4	5	5	7	No	No	No	No	Yes
Ryan1	4	5	3	No	No	No	No	Yes
Ryan2	5	5	3	No	No	No	No	Yes
Ryan3	3	4	5	No	No	No	Yes	Yes
Ryan4	2	3	5	No	No	No	Yes	Yes
Stacey1	6	5	2	No	Yes	No	No	No
Stacey2	6	5	2	No	Yes	No	No	No
Stacey3	6	6	2	No	Yes	No	No	No
Stacey4	4	5	4	No	Yes	No	No	No

Table 3. Questions used for qualitative evaluation

Character	Questions-nbr	Question
Janine	Q1	What is causing Ryan to be sick in bed?
Janine	Q2	How does Janine attempt to kill Ryan while in the hospital?
Janine	Q3	What happens when Janine attempts to play recording of Stacey?
Janine	Q4	Who stabbed Janine?
Janine	Q5	**Who gives Janine the recording of Stacey?**
Ryan	Q1	How does Janine attempt to kill Ryan in the hospital?
Ryan	Q2	What does Ryan do when Janine is lying in the hospital?
Ryan	Q3	Where is Ryan trapped?
Ryan	Q4	**What does Ryan tell Phil he can do for him?**
Ryan	Q5	**Who is Ryan with when going to put his name on the babies birth cert?**
Stacey	Q1	Who climbs up the roof to talk Stacey out of jumping off?
Stacey	Q2	**What does Stacey reveal when in a cell with Janine, Kat, and Pat?**
Stacey	Q3	What does Stacey admit to her mum in bedroom when mum is upset?
Stacey	Q4	Who confronts Stacey in restroom where Stacey finally admits to killing Archie?
Stacey	Q5	Who calls to Stacey's door to tell her to get her stuff and go after Stacey's mum had called the police?

One of the key contribution of this paper is to have demonstrated that despite some noise from face detection and recognition, this method enables to capture multiple important plot points for all three query characters. We also conclude that adding more shots to the summaries did, quite surprisingly, not always allow to answer more key moments related questions. Finally, we would like to pinpoint the fact that the task of choosing important sequences that would answer unknown questions, is very challenging for humans. Indeed, when generating the runs, having read the summaries but not having watched the videos, we find it challenging to decide which sequences should be included in the summary. It would be interesting to know how much the score would improve if we would know the questions before evaluation.

REFERENCES

George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. 2020. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of the TRECVID 2020 Workshop*. NIST, Gaithersburg, MD, USA.

Mats Sjöberg, Hamed R. Tavakoli, Zhicun Xu, Héctor Laria Mantecón, and Jorma Laaksonen. 2018. PicSOM Experiments in TRECVID 2018. In *Proceedings of the TRECVID 2018 Workshop*. NIST, Gaithersburg, MD, USA.