GLOBOPLAY THUMBNAILS: USING AI TO EXTRACT THE BEST FRAME TO REPRESENT DRAMA SERIES

Edmundo Hoyle, Álvaro Antelo, Igor Coutinho

Researchers at GLOBO, Brazil

ABSTRACT

Traditionally, the process of choosing the best thumbnail is having human curators or editors select it. It requires technical expertise as well as a good notion of which specific target population that content should be directed to. We present an algorithm that automatically extracts the “Best Frame" from the drama series’ episodes to be the thumbnail using not only the video but also the episode summary text and closed captions (CC). This “Best Frame” carries the essential characteristics that human professionals would look for while trying to manually select a thumbnail.

INTRODUCTION

Globoplay is an Over-the-top platform developed by GLOBO that offers access to broadcast content directly over the internet: streaming live content, drama series, documentaries, news and entertainment programs. Each video published in Globoplay is accompanied by a thumbnail and a short summary. In the case of drama thumbnails, to increase engagement, it is common to picture a relevant event that happened on that particular content. Globoplay’s interface can be seen as an example in Figure 1.

Like in other digital video platforms, it takes special care in the way products are presented to the users in order to increase their interest and engagement. In this scenario, thumbnails play a very important role in online video display. As the most representative snapshot, they are supposed to capture the essence of each video and provide an accurate first impression to the viewers. Studies suggest that people look at thumbnails extensively when browsing online videos [1]. An accurate thumbnail would ultimately make a video more attractive to watch, which, in turn, leads to an increase in ad revenue [2].

This work introduces a novel algorithm responsible for the automatic thumbnail selection of drama series content published on Globoplay. The presented method uses not only image features but also text metadata to make the best thumbnail choice. It aims to achieve the same level of quality obtained by the manual process.
In the next section the problem context will be detailed further; following, the process of automation and its algorithm is explained; after that, some analysis results are shown; and finally, the work conclusions and future works are presented.

Figure 1 – Globoplay screenshots. From left to right, Home Screen, Categories, list of Drama Series, thumbnail and summary from an episode.

PROBLEM CONTEXT: MANUAL THUMBNAIL SELECTION

At Globo, manual thumbnail selection is done by professionals following a rigorous process involving a set of rules and criteria in order to find a particular frame that best represents the entire video. That process is usually time consuming. Globo regularly produces five different drama series at the same time on a daily basis; automating the thumbnail selection process is, therefore, a desirable goal due to a much higher productivity potential.

The drama series produced by Globo are available in Globoplay only after they are exhibited on live public TV. During the automatic process of video publishing, for safety, the thumbnail is set to a fixed frame at 5 seconds of content. Afterwards, the thumbnail for each content is edited manually. That work not only takes time but is also time sensitive since the editors need all of that content to be displayed with an accurate thumbnail.

The process of choosing the perfect frame for thumbnail consists of some rule-checking. A thumbnail has the following restrictions:

1. It must contain relevant events that happened in the chapter.
2. It should be clear, without blur.
3. Can't include nudity.
4. Can't show weapons.
5. It must always contain people.

Apart from those rules, some other parameters are considered. For example, characters shouldn't be making grimaces, full shots and motion shots are also avoided.

AUTOMATION PROCESS: AUTOMATIC THUMBNAIL SELECTION

The complete workflow can be broken in three processes. First, smaller relevant video segments are extracted from the episode using text metadata. After that, two frames from each segment are chosen as candidates based on aesthetic metrics. Finally, the resulting list is filtered using the restrictions imposed by the editors, which were listed previously

Process 1: Finding Relevant Video Segments

Each video of an episode has its own text metadata, a summary of the content, usually containing relevant events from the episode. The first step to automatically choose a thumbnail to display is to find segments within the video that best match these events. For that, a text search engine technique is used.

With a bag-of-words or bag-of-ngrams paradigm, where text is represented uniquely by its terms, search engines use a measure of relevance of each term in a document. This measuring can be made in various ways. It can be based on binary word presence, word counting or an indexing known as TF-IDF [3]. TF-IDF values term frequency in a document positively and the presence of the same term across all documents negatively, creating a balance between term frequency and specificity. Having text documents represented by its terms' relevance allows these documents to be compared with other documents for similarities. Such approach has also been used for video recommendation in Globoplay [3].

Starting the process, the text metadata is split into sentences. Each of those sentences is compared with all closed caption lines. This is analogous to using a search engine to look for the text sentence inside the closed caption documents(lines). That comparison creates a rank of similarity between text sentence and closed caption lines.

In other words, the search uses a TF-IDF indexing to rank the similarities between each closed caption line and each text metadata sentence. The most similar closed caption line to each sentence defines the timecode input and timecode output for a relevant video segment. This process outputs a list of relevant video segments, one for each sentence in the text metadata of the content. Figure 2 illustrates this process.

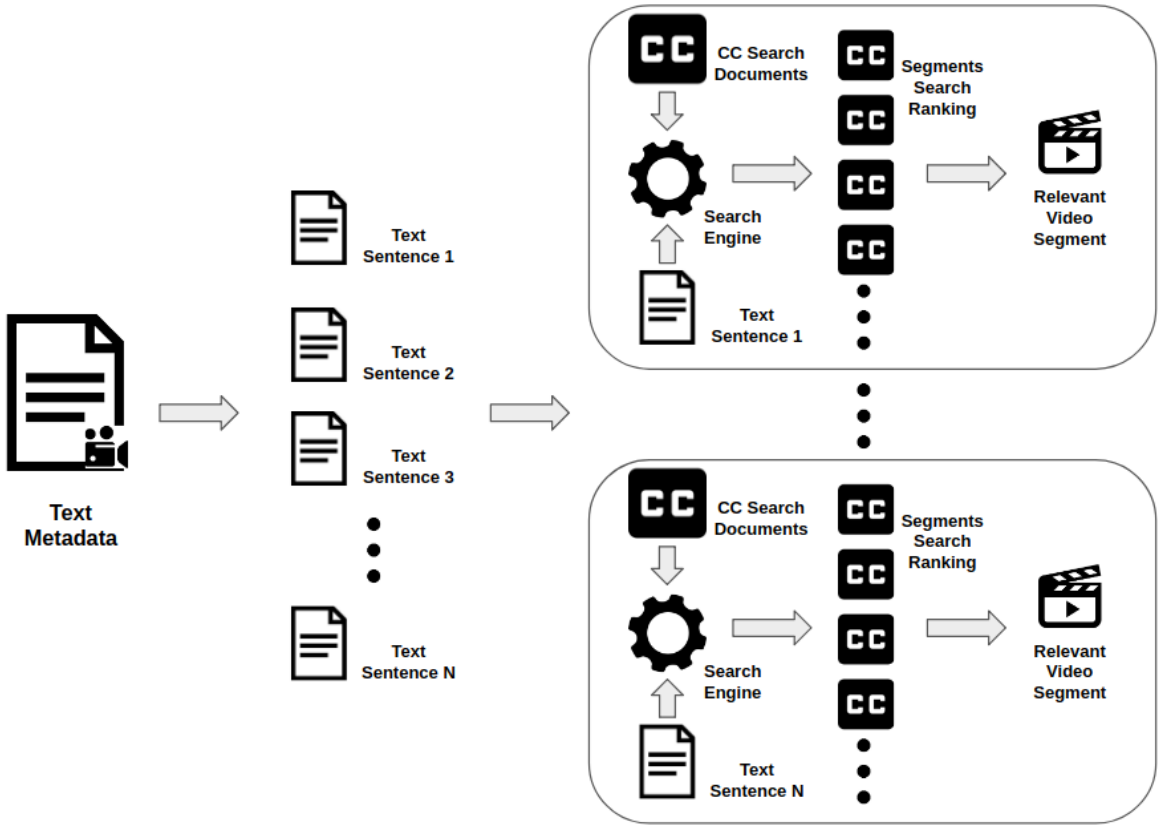

Figure 2 – Using text to find relevant events in the video.

Process 2: Select the Best-Looking Frames in the Segment

When searching for the best-looking frame inside a video segment, at first, a summarization is made, sampling one of every ten frames from the segment.

With the resulting frames in hand, the next step is to classify and rank them according to their aesthetics. For that, a supervised learning method was used with a dataset using the previous episodes of the drama series. The positive label was given to the sample frames that were in fact chosen as thumbnails manually. Some image metrics were used as features of that classifier such as color distribution features, texture and quality features. Those metrics are also used in previous works in thumbnail classification, such as [5].

With the classifier trained, the frames are ranked according to their probability of being part of the positive class (chosen as thumbnail). For each video segment, this process suggests the first and second frame from that rank.

Process 3: Applying restrictions using Deep Learning

After executing the previous step, a list of candidate frames is ready. That list has already been checked for event relevance and frame aesthetics. The next step is to guarantee that all images suggested respect the restrictions imposed by the professional editors.

To detect people, nudity and weapons in the imagens we used Amazon Rekognition. This service brings numerous image analysis features that could be easily integrated in the system. With Rekognition, the system can detect particular objects (like weapons), scenes, and faces in images. Therefore, it could detect explicit and suggestive adult content in the candidate frames and rule them out.

The scheme in Figure 3 summarizes the whole system workflow.

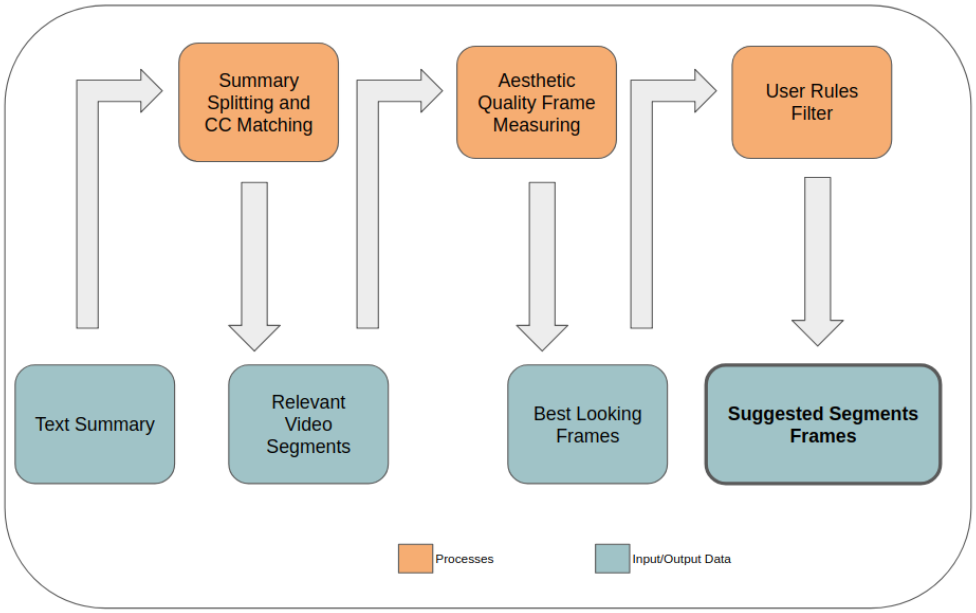

Figure 3 – The complete workflow.

RESULTS & ANALYSIS

Before the algorithm was released, it was tested using old episodes from the drama series "*Malhação*". The intention was to verify if the manually chosen thumbnail would in fact be chosen by the algorithm. The results were promising and showed that in most cases (nearly 90%) one of the images suggested by the algorithm was in the same video segment where the manual thumbnail was chosen. Figure 4 shows some samples.

The system was integrated with Globoplay on January 23th 2019. The following results and analysis were based in the automatic thumbnail selection for episodes of the drama series "*Malhação*" until May 3rd. In total, the thumbnail of 73 episodes was suggested by the system. In 92% (67 episodes) of all cases, the editors accepted one of the thumbnails presented as relevant to that episode. In roughly one third of those cases (21 episodes) the editor agreed that the first suggested image was in fact the best choice of thumbnail for that episode. Only 8% (6 episodes) had their suggestion list completely rejected. The chart in Figure 5 illustrates these results.

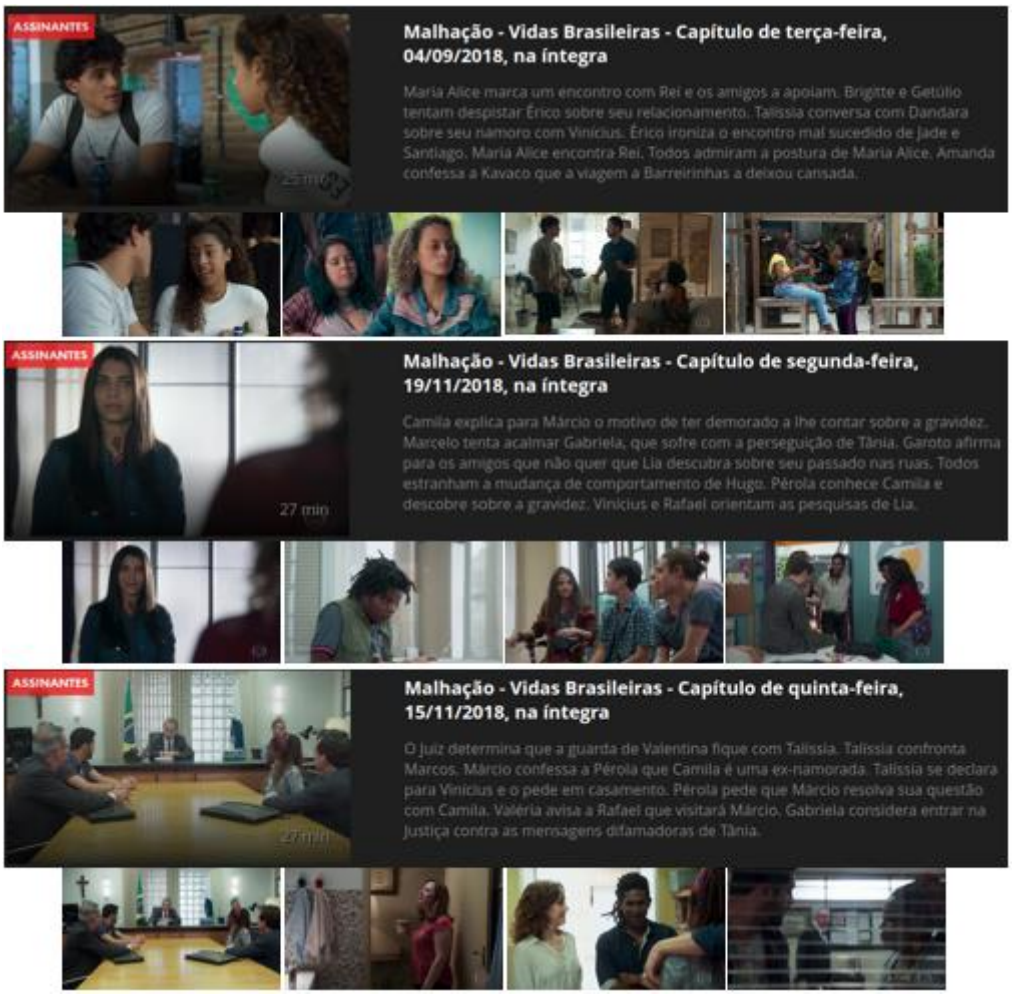

Figure 4 – Examples of manually selected thumbnails versus the automatic suggestion

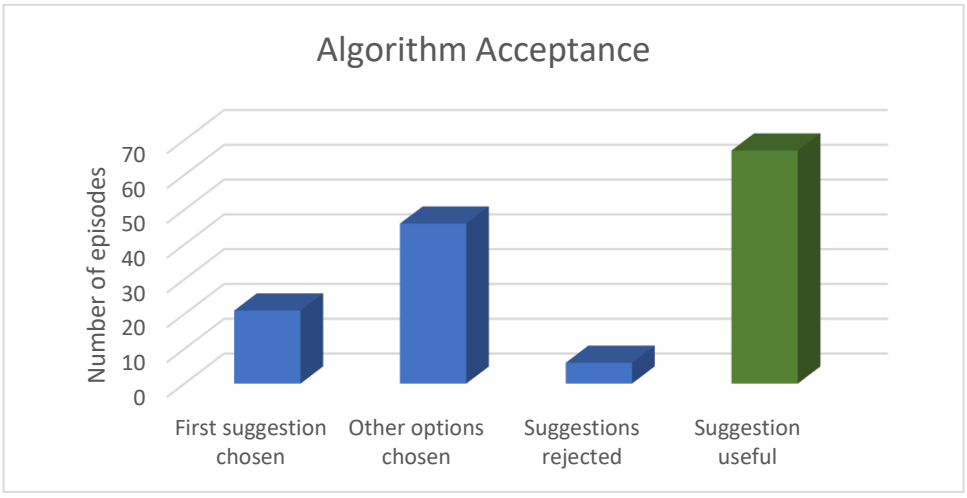

Figure 5 –The successful suggestions in green and all other partials in blue

Figure 6 – Images showing samples when the professionals didn't accept the first thumbnail suggested but they used other similar candidates.

Some of the reasons presented by the editors for eventually needing to intervene in the algorithm choice were:

1. An image where the character was facing forward was preferred.
2. The facial expression of the character was not ideal.
3. The character was gazing down or in a not ideal direction.
4. Although the thumbnail was a good choice the editor preferred a different event.
5. The characters were far in the background and hard to recognize.

These cases are exemplified in Figure 6, sorted respectively. Most of those reasons for intervention are highly subjective. Although they could be further addressed, subjectivity is something hard to model in an algorithm. In spite of that, the vast majority of the suggestions successfully captured relevant events in the video and frames within those events.

CONCLUSIONS

We presented an automatic method to extract the most representative or "Best Frame" from drama series' episode. The method proposed uses text metadata to find relevant events in the video. From there, the best frames are chosen using aesthetic metrics and filtered by user defined rules. This method was introduced as a feature in the Globoplay platform to deliver the "best thumbnails" and some alternatives for the drama series' episode of

"*Malhação*". Conclusively, it obtained great results with 92% of accurate thumbnails being suggested and accepted by professional editors. Furthermore, it greatly increased productivity since in most cases the editors only had to verify the thumbnail choice instead of actually looking for it. This will give the editors more time and opportunity to focus on more important tasks in content production.

FUTURE WORKS

There are a few next steps for this project. First, the system should be used to cover all Globo drama series. Second, some A/B testing needs to be done to improve result metrics. This testing could eventually verify if the rejected thumbnail suggestions would influence negatively in the video consumption.

Finally the subjective reasons for editor intervention need to be addressed to make the system even more accurate and improve productivity further.

REFERENCES

[1] B. Georg, C. Edward e M. M. Ringel, “What Do You See when You'Re Surfing?: Using Eye Tracking to Predict Salient Regions of Web Pages,” em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2009.

[2] D. H. a. J. A. J. Jiang, "Searching, browsing, and clicking in a search session: changes in user behavior by task and over time.," in *SIGIR*, 2014.

[3] S. Qaiser e R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” em *International Journal of Computer Application*, New York, USA, 2018.

[4] M. Souza, J. Castellani, D. Monteiro and C. Queiroz, " Big Data For Data Journalism, Enhanced Business Analytics And Video Recommendation at Globo," in *2017 IBC Conference*, Amsterdam, Netherlands, 2017.

[5] Y. Song, M. Redi, J. Vallmitjana e A. Jaimes, “To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos,” em *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, New York, NY, USA, 2016.

[6] W. v. d. Meer, K.-K. R. Choo, N.-A. Le-Khac and M.-T. Kechad, "Investigation and Automating Extraction of Thumbnails Produced by Image viewers," in *IEEE Trustcom/BigDataSE/ICESS*, Sydney, NSW, Australia, 2017.

[7] J. Huang, H. Chen, B. Wang and S. Lin, "Automatic Thumbnail Generation Based on Visual Representativeness and Foreground Recognizability.," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.