Book2Movie: Aligning Video scenes with Book chapters

Makarand Tapaswi, Martin Bäuml, Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
https://cvhci.anthropomatik.kit.edu/projects/mma

High grossing films and TV series are often adapted from novels [1]. Motivated by the increase in adaptations, we introduce a new problem: to align parts of the original source novel to the video adaptation. To the best of our knowledge, we are the first to jointly analyze and align books and films.

A good alignment opens multiple applications and can be used to understand videos at the story level. A direct application is to find differences between the novel and the film story line, while, another is to use the rich descriptions in the novel as weak labels to improve video understanding and automatic description. An alternative application is story-based video retrieval. Previously, [3] aligns sentences of the plot synopses with video shots and uses this text to bridge the gap between the query and video for video retrieval. Books can augment these synopses as they are much larger and thus provide a higher chance of a query matching the text source.

As a first work in this direction, we restrict our granularity to *video scenes* and *book chapters*. To emphasize on finding differences, we not only predict which scene belongs to which chapter, but also indicate when a scene is not part of the book.

Previous work on text-to-video alignments Closest to our goal and approach are two works which use dynamic programming. [2] aligns video shots to transcripts without using subtitles; and [3] aligns plot synopsis sentences with shots of the video. However, both proposed models make strong assumptions about the monotonicity of the alignment, *i.e.*, they assume that shots and the text appear sequentially, and that every shot is assigned to some sentence. This is often not true for novel-to-film adaptations of complex story lines. We address this problem in this paper.

Data set We introduce a new data set consisting of two popular adaptations. (i) GOT: features season 1 of the TV series *Game of Thrones*, and its corresponding novel from *A Song of Ice and Fire*; and (ii) HP: consists of the first film and novel in the series – *Harry Potter and the Sorcerer's Stone*. Our choice is motivated by the large differences in the two adaptations.

GOT includes multiple concurrent story lines taking place at different locations in the world and thus favors intertwined adaptations. GOT also has a large cast list (90+) with many of them being primary characters and the book and the video material are 4 times larger than HP. On the other hand, HP is centered around a few characters, and has a linear story line.

Scene-to-chapter alignment Our goal is to assign for each scene the best possible chapter which maximizes a measure of similarity between them, subject to story progression constraints. We use two cues to compute the similarity between scenes and chapters. Characters play a very important role in any story. We use name mentions in the book and face track identification in the video and normalize the scores to construct a character-based similarity. Our second cue is obtained by matching dialogs between the book and the video. The matching is performed by scoring the identical words in the longest common subsequence between the dialogs.

We model the alignment problem as finding the *shortest path* in a grid-structured graph (green nodes) with fully connected edges between consecutive columns (see Fig. 1). Every valid path corresponds to a possible alignment. The edge weights depend on the similarity between the corresponding chapter and the scene. We set up prior distances between the edges of the graph to conform to a linear spread (from top-left to bottom-right). Our similarity cues are used to reduce the incoming edge distances for a node which encourage the path to go through high scoring similarity nodes. Finally, an additional layer of nodes (blue) is used to allow assignment of scenes to a null chapter. The graph based representation provides a fast and efficient solution while circumventing challenges of monotonicity and allowing scenes to be not assigned to any chapter.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

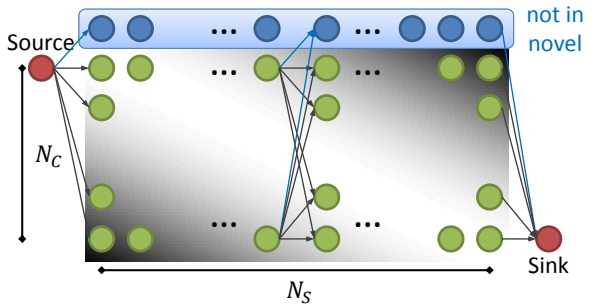

Figure 1: Illustration of the graph used for alignment.

Experiments Our ground truth annotation is at the shot level and consists of the set of shots which belong to each chapter. We evaluate the alignment performance using accuracy, the fraction of shots assigned to the correct chapter. We also consider metrics to evaluate detection of shots which are not part of the book. On GOT, our method is able to obtain an accuracy of 75.7% while the closest baseline, an improvement over [3] yields 60.7%.

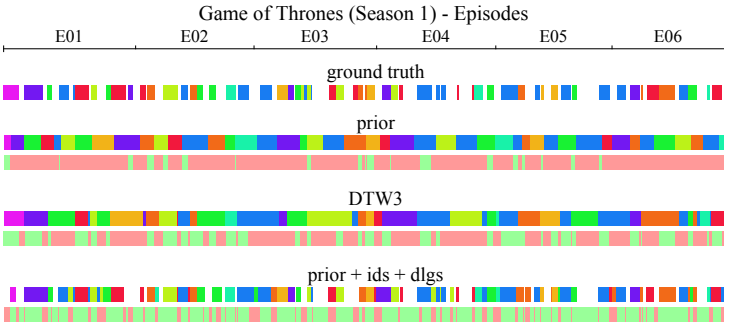

Figure 2: Alignment performance on GOT for various methods.

Fig. 2 shows the alignment performance of the various methods. The ground truth alignment is row 1. Chapters are indicated by colors, and white spaces imply that those shots are not part of the novel. For each subsequent method, the color of the first sub-row indicates the chapter to which every scene is aligned. Comparing vertically against the ground truth we can determine alignment accuracy. For simplicity, the second sub-row of each method indicates whether the alignment is correct (green) or wrong (red).

Rich descriptions We present in Fig. 3 samples of rich descriptions obtained from the novel and their corresponding video shots which are obtained pseudo-automatically by looking in a small interval.

(a) GOT: (Ch3, P8, E01 19:50) At the center of the grove an ancient weirwood brooded over a small pool where the waters were black and cold.

(b) HP: (Ch10, P78-79, M 1:06:59) Hermione rolled up the sleeves of her gown, flicked her wand, and said, "Wingardium Leviosa!" Their feather rose off the desk and hovered about four feet above their heads.

Figure 3: Describing video shots using passages from the novel.

[1] Where do highest-grossing screenplays come from? http://stephenfollows.com, Jan. 2014. Retrieved 2014-11-14.

[2] P. Sankar, C. V. Jawahar, and A. Zisserman. Subtitle-free Movie to Script Alignment. In *BMVC*, 2009.

[3] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based Video Retrieval in TV series using Plot Synopses. In *ACM ICMR*, 2014.