Book2Movie: Aligning Video scenes with Book chapters
Supplementary material

Makarand Tapaswi Martin Bäuml Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{makarand.tapaswi, baeuml, rainer.stiefelhagen}@kit.edu
http://cvhci.anthropomatik.kit.edu/projects/mma

1. Introduction

In addition to the paper, this supplementary material provides more insights into the novel-to-film alignment and the results. We specially focus on three aspects of the paper.

Firstly, Sec. 2 presents a detailed qualitative analysis of the alignment methods, and also visualizes the ground truth and predicted alignment for our second data source HP (*Harry Potter and the Sorcerer's Stone*). We further motivate the problem in Sec. 3 and present a list of interesting articles and visualizations that analyze the relation between novels and their film adaptations. Finally, Sec. 4 presents many more examples of describing video shots using passages from the source novel.

Late 2014 saw a sharp rise in interest to jointly model images and text, especially to generate automatic descriptions of images [10, 11, 12, 13]. Convolution Neural Network (CNN) models have gained popularity and are often used for representing the image, while Recurrent Neural Networks (RNNs) are used to model and generate the textual descriptions.

Videos are typically much more challenging than single images, however automatically describing videos would facilitate their indexing and thus allow to search through this rapidly increasing data form. Towards this goal, and especially for high-level understanding of TV series and films, alignment of video scenes to novels can provide a large amount of weak labels for training such models.

2. Qualitative analysis of alignments

Similar to Fig. 1 and Fig. 4 of the original paper which present the ground truth and predicted alignment for GOT (*Game of Thrones*), we present here the results for the other data source HP.

Ground truth alignment for *Harry Potter* Fig. 3 presents the ground truth alignment for our second data source *Harry Potter* (HP). Compare this against Fig. 1 for the *Game of Thrones* (GOT) data set (repeated here from Fig. 1 of the paper). HP is arguably a much simple adaptation and the film stays true to the novel. The alignment is also linear and chapters and video parts do not cross. On the other hand, GOT involves a number of shots which do not belong to parts of the novel, and involves complicated intertwining between various story lines.

Predicted alignment for HP, and comparison to GOT Fig. 2 (repeated here from Fig. 4 of the paper) and Fig. 4 visualize the assignment of shots to book chapters using three methods for GOT and HP respectively.

For GOT (Fig. 2), we see that our method (*prior* + *ids* + *dlgs*) is able to mimic the ground truth alignment best. In contrast, note how both *prior* and *DTW3* assign every scene to book chapters. While *DTW3* makes correct use of cues, it has two major flaws: (i) each chapter can only be assigned a contiguous set of scenes; and (ii) chapter c cannot be assigned before chapter $c-1$.

For HP (Fig. 4), firstly we see a large gap at the end of the video. This corresponds to the long credits sequence that appears at the end of the film, and forms a large (9 minutes of the 2h 32m film) scene.

The *prior* method fails to consider that the initial chapters of the novel are shortened in the video representation resulting in many alignment errors. In contrast to GOT, *DTW3* shows slight improvement as compared to our proposed approach *prior* + *ids* + *dlgs*. This is due to an inherent property of our proposed approach that requires evidence to be found to associate a chapter with a scene. Fig. 4 clearly shows how our alignment allows jumping between chapters, while DTW3 does not.

Failure analysis of face identification As face track identities are an important cue for our alignment, we analyze the impact of errors in identification on the alignment performance. Using ground truth face identities provides an alignment accuracy of 77.6 for GOT (75.7 with automatic id) and 90.3 for HP (89.9 with automatic id). Only a small improvement is seen since the id errors get averaged when considering all tracks in a scene.

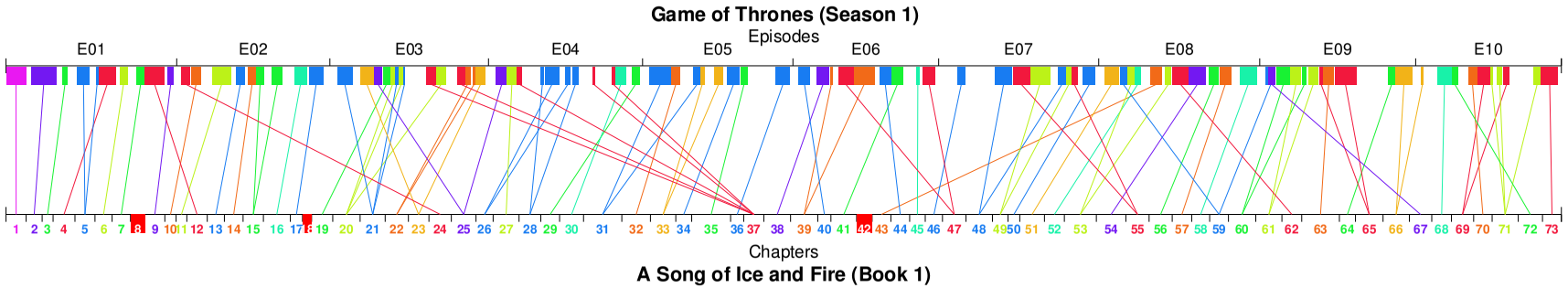

Figure 1. This figure is best viewed in color. We present the ground truth alignment of chapters to video for our first data set *Game of Thrones* (GOT). This figure is repeated here for ease of comparison and appears as Fig. 1 in the paper. TV series episodes are depicted at the top, and book chapters at the bottom. The alignment is indicated by drawing a line between book chapters to parts of the video.

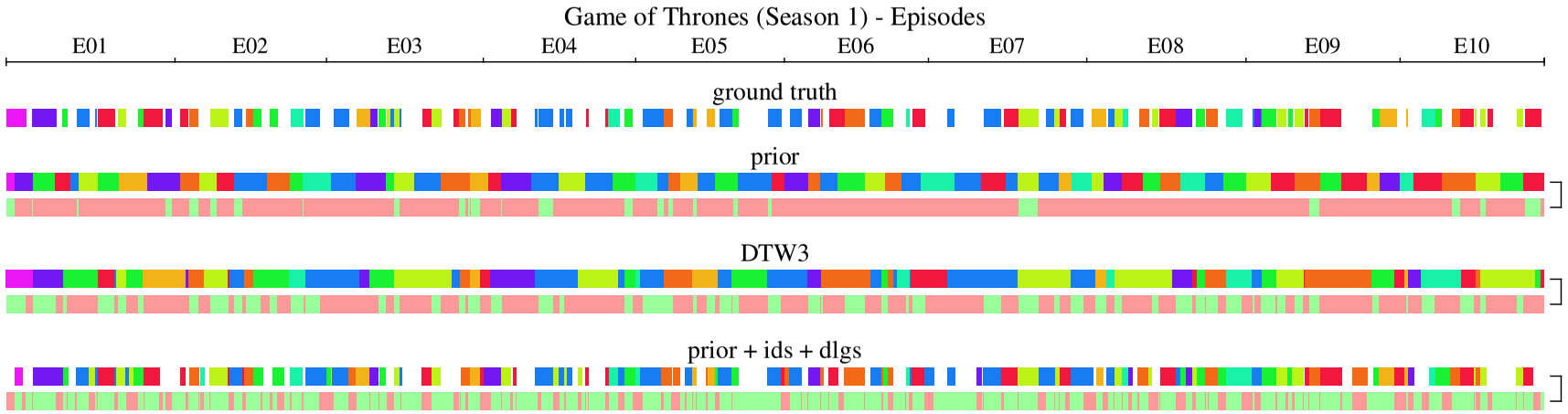

Figure 2. This figure is best viewed in color. We visualize the alignment as obtained from the various methods. This figure is repeated here for ease of comparison and appears as Fig. 4 in the paper. The ground truth alignment (row 1) is the same as presented in Fig. 1. Chapters are indicated by colors, and empty areas (white spaces) indicate that those shot are not part of the novel. We present alignment performance of three methods: *prior* (row 2), *DTW3*(row 3) and our approach *prior* + *ids* + *dlgs* (row 4). The color of the first sub-row for each method indicates the chapter to which every scene is assigned. Comparing vertically against the ground truth we can determine the accuracy of the alignment. For simplicity, the second sub-row of each method indicates whether the alignment is correct (green) or wrong (red).

3. Novels and films – interesting links

In this section we present some articles that might interest the reader and help motivate the problem of co-analyzing novels and their video adaptations.

An interesting analysis shows that a large number of highest-grossing screenplays come from literary adaptations [9]. This is a very large and yet untapped resource for joint analysis.

Video adaptations have often resulted in boosted sales for the original novels. Consider the articles related to popular adaptations such as *Harry Potter* series [3], *Game of Thrones* series [7], *Hunger Games* trilogy [6] and even very old novels from authors like *Charles Dickens* [4].

Visualization [1, 5] of differences between books and films or listing differences [2] is very popular among the fans of the stories. For the *Game of Thrones* TV series, there is even interest in finding which episodes correspond to which book chapters [8]. Note that we go deeper by dividing the episode into multiple scenes for the alignment.

4. Rich descriptions

Similar to Fig. 5 from the paper, we now present many more examples of selected video frames and their descriptions. The examples of Fig. 5 are from the *Harry Potter* data set and those from Fig. 6 and Fig. 7 are from *Game of Thrones*.

We believe that such descriptions can be used as an initial step towards automatically understanding and describing the story conveyed in TV series and films.

References

[1] 32 Differences between Children's Books and Their Movies. http://visual.ly/32-differences-between-books-and-their-movies. Retrieved 2014-11-21.

[2] That was Not in the Book. http://thatwasnotinthebook.com/. Retrieved 2014-11-21.

[3] 'Deathly Hallows' film breathes life into Harry Potter book sales. http://www.nielsen.com/us/en/insights/news/2010/deathly-hallows-film-breathes-life-into-harry-potter-book-sales.html, Nov. 2010. Retrieved 2014-11-21.

[4] Dickens book sales boosted by television adaptations. http://www.bbc.com/news/entertainment-arts-16579263, Jan. 2012. Retrieved 2014-11-21.

[5] "Harry Potter" Characters in the Books vs. the Movies. http://www.buzzfeed.com/keenan/harry-potter-characters-in-the-books-vs-the-mov, Jul. 2012. Retrieved 2014-11-21.

[6] 'Hunger Games' Movie Fuels Sharp Rise in Book Sales. http://www.hollywoodreporter.com/news/hunger-games-twlight-book-sales-versus-jennifer-lawrence-josh-hutcherson-305457, Mar. 2012. Retrieved 2014-11-21.

[7] Adaptation: How A Song of Ice and Fire Book Sales Move with Game of Thrones. http://nielsentopten.com/2014/04/18/adaptation-how-a-song-of-ice-and-fire-book-sales-move-with-game-of-thrones/, Apr. 2014. Retrieved 2014-11-21.

[8] How does the Game of Thrones Series Line Up With the Books? http://www.slate.com/articles/arts/television/2014/04/game_of_thrones_hbo_which_episodes_portray_which_chapters_from_a_song_of.html, Apr. 2014. Retrieved 2014-11-21.

[9] Where do highest-grossing screenplays come from? http://stephenfollows.com/where-do-highest-grossing-screenplays-come-from/, Jan. 2014. Retrieved 2014-11-21.

[10] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015.

[11] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *Transactions on Association for Computational Linguistics*, 2015.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. In *NIPS Deep Learning workshop*, 2014.

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015.

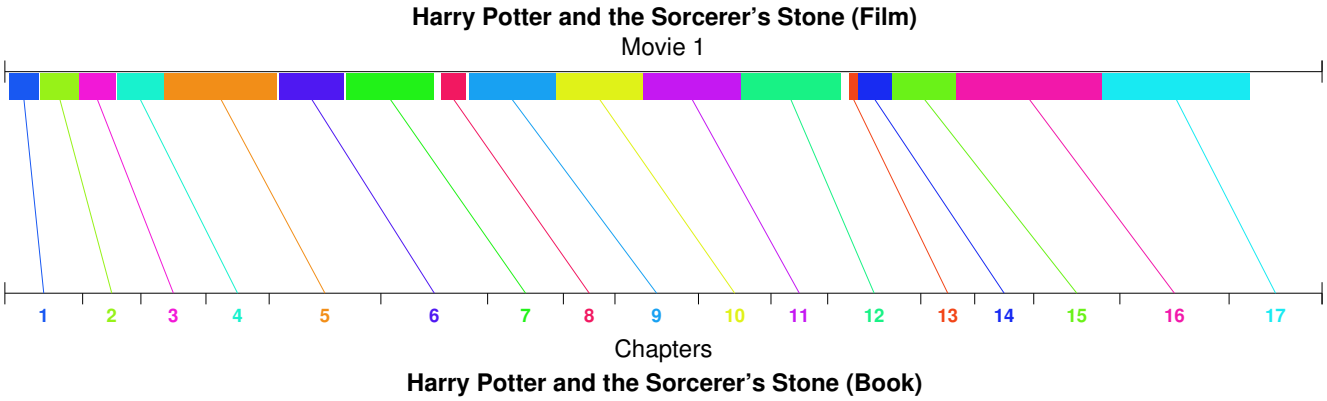

Figure 3. This figure is best viewed in color. We present the ground truth alignment of chapters to video for our second data set *Harry Potter* (HP). The movie scenes are depicted at the top, and book chapters at the bottom. The alignment is indicated by drawing a line between book chapters to parts of the video. Compare the simplicity of this adaptation against GOT, Fig. 1.

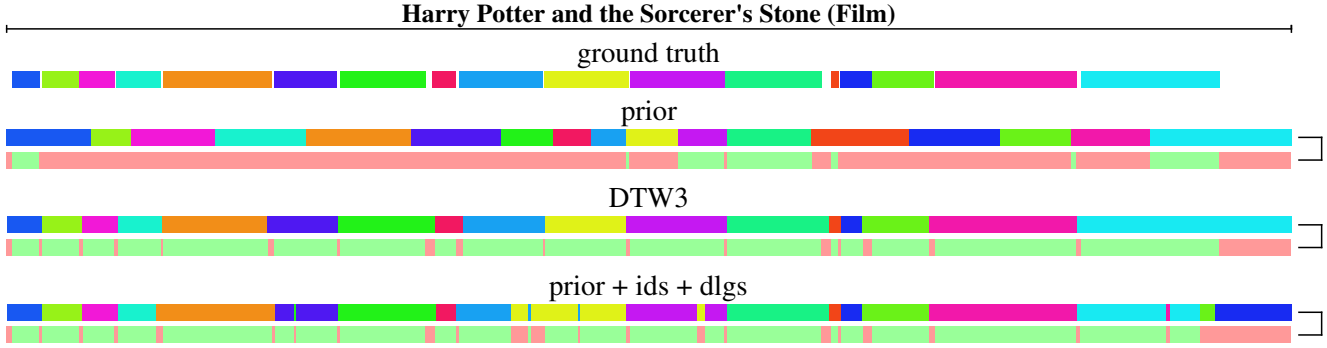

Figure 4. This figure is best viewed in color. We visualize the alignment as obtained from the various methods. The ground truth alignment (row 1) is the same as presented in Fig. 3. Chapters are indicated by colors, and empty areas (white spaces) indicate that those shot are not part of the novel. We present alignment performance of three methods: *prior* (row 2), *DTW3*(row 3) and our approach *prior + ids + dlgs* (row 4). The color of the first sub-row for each method indicates the chapter to which every scene is assigned. Comparing vertically against the ground truth we can determine the accuracy of the alignment. For simplicity, the second sub-row of each method indicates whether the alignment is correct (green) or wrong (red).

(a) (Ch14, P66, M 1h40m30s): All at once there was a scraping noise and the egg split open. The baby dragon flopped onto the table. It was n't exactly pretty; Harry thought it looked like a crumpled, black umbrella. Its spiny wings were huge compared to its skinny jet body, it had a long snout with wide nostrils, the stubs of horns and bulging, orange eyes.

(b) (Ch1, P90, M 2m14s): A low rumbling sound had broken the silence around them. It grew steadily louder as they looked up and down the street for some sign of a headlight; it swelled to a roar as they both looked up at the sky – and a huge motorcycle fell out of the air and landed on the road in front of them.

(c) (Ch10, P132, M 1h11m37s): The club flew suddenly out of the troll's hand, rose high, high up into the air, turned slowly over – and dropped, with a sickening crack, onto its owner's head. The troll swayed on the spot and then fell flat on its face, with a thud that made the whole room tremble.

(d) (Ch3, P89, M 10m50s): On Sunday morning, Uncle Vernon sat down at the breakfast table looking tired and rather ill, but happy.

(e) (Ch6, P271, M 38m23s): Hagrid's big hairy face beamed over the sea of heads. "C'mon, follow me – any more firs' years? Mind yer step, now! Firs' years follow me!"

(f) (Ch7, P131, M 46m55s): Harry, who was starting to feel warm and sleepy, looked up at the High Table again. Professor Quirrell, in his absurd turban, was talking to a teacher with greasy black hair, a hooked nose, and sallow skin.

(g) (Ch17, P7, M 2h07m39s): "Severus?" Quirrell laughed, and it wasn't his usual quivering treble, either, but cold and sharp. "Yes, Severus does seem the type, doesn't he? So useful to have him swooping around like an overgrown bat. Next to him, who would suspect p-p-poor, st-stuttering P-Professor Quirrell?"

(h) (Ch3, P34, M 9m07s): Harry went back to the kitchen, still staring at his letter. He handed Uncle Vernon the bill and the postcard, sat down, and slowly began to open the yellow envelope.

(i) (Ch2, P77, M 6m40s): The snake suddenly opened its beady eyes. Slowly, very slowly, it raised its head until its eyes were on a level with Harry's.

Figure 5. This figure is best viewed on screen. We present more examples of rich attributes (highlighted) that are easily extracted from the *Harry Potter* data set. The location of the description is abbreviated as *chapter number* (Ch), *paragraph number* (P) and the location in the correct video as movie M in hours, minutes and seconds. The caption text is directly used from the novel. The various attributes are highlighted in the caption below every frame.

(a) (Ch32, P29, E05 10m06s): The rain had finally stopped and dawn light was seeping through the wet cloth over his eyes when Catelyn Stark gave the command to dismount. Rough hands pulled him down from his horse, untied his wrists, and yanked the hood off his head. When he saw the narrow stony road, the foothills rising high and wild all around them, and the jagged snowcapped peaks on the distant horizon, all the hope went out of him in a rush.

(b) (Ch23, P83, E03 50m37s): Three days later, at midday, her father's steward Vayon Poole sent Arya to the Small Hall. The trestle tables had been dismantled and the benches shoved against the walls. The hall seemed empty, until an unfamiliar voice said, "You are late, boy." A slight man with a bald head and a great beak of a nose stepped out of the shadows, holding a pair of slender wooden swords. "Tomorrow you will be here at midday," He had an accent, the lilt of the Free Cities, Braavos perhaps, or Myr.

(c) (Ch28, P66, E04 29m45s): The master called over a tall lad about Robb's age, his arms and chest corded with muscle. Thick hair, shaggy and unkempt and black as ink. The shadow of a new beard darkened his jaw. "This is Gendry. Strong for his age, and he works hard. Show the Hand that helmet you made, lad." Almost shyly, the boy led them to his bench, and a steel helm shaped like a bull's head, with two great curving horns.

(d) (Ch23, P43, E03 7m2s) Ned was aghast. "Aerys Targaryen left a treasury flowing with gold. How could you let this happen?" Littlefinger gave a shrug. "The master of coin finds the money. The king and the Hand spend it."

(e) (Ch3, P8, E01 19m50s) At the center of the grove an ancient weirwood brooded over a small pool where the waters were black and cold. "The heart tree," Ned called it. The weirwood's bark was white as bone, its leaves dark red, like a thousand bloodstained hands. A face had been carved in the trunk of the great tree, its features long and melancholy, the deep-cut eyes red with dried sap and strangely watchful.

(f) (Ch33, P58, E05 25m19s) She was miles from the castle, but from anywhere in King's Landing you needed only to look up to see the Red Keep high on Aegon's Hill, so there was no danger of losing her way. Her clothes were almost dry by the time she reached the gatehouse. The portcullis was down and the gates barred, so she turned aside to a postern door. The gold cloaks who had the watch sneered when she told them to let her in. "Off with you," one said.

Figure 6. This figure is best viewed on screen. We present more examples of rich attributes (highlighted) that are extracted from the *Game of Thrones* data set. The location of the description is abbreviated as *chapter number* (Ch), *paragraph number* (P) and the location in the correct video as *episode number* (E) and minutes and seconds. The caption text is directly used from the novel. The various attributes are highlighted in the caption below every frame.

(a) (Ch49, P69, E07 45m02s): They said the words together, as the last light faded in the west and gray day became black night. "Hear my words, and bear witness to my vow," they recited, their voices filling the twilit grove. "Night gathers, and now my watch begins. It shall not end until my death. I shall take no wife, hold no lands, father no children." ...

(b) (Ch65, P129, E09 32m40s): It was enough. Ser Jorah brought his longsword down with all the strength left him, through flesh and muscle and bone, and Qotho's forearm dangled loose, flopping on a thin cord of skin and sinew. The knight's next cut was at the Dothraki's ear, so savage that Qotho's face seemed almost to explode.

(c) (Ch29, P74, E04 52m16s): She did not know what was more satisfying: the sound of a dozen swords drawn as one or the look on Tyrion Lannister's face.

(d) (Ch17, P52, E02 51m09s): He left the room with his eyes burning and his daughter's wails echoing in his ears, and found the direwolf pup where they chained her. Ned sat beside her for a while. "Lady," he said, tasting the name. He had never paid much attention to the names the children had picked, but looking at her now, he knew that Sansa had chosen well. She was the smallest of the litter, the prettiest, the most gentle and trusting. She looked at him with bright golden eyes, and he ruffled her thick gray fur.

(e) (Ch68, P20, E10 10m50s): Sansa stared at him, seeing him for the first time. He was wearing a padded crimson doublet patterned with lions and a cloth-of-gold cape with a high collar that framed his face. She wondered how she could ever have thought him handsome. His lips were as soft and red as the worms you found after a rain, and his eyes were vain and cruel. "I hate you," she whispered.

(f) (Ch38, P76, E06 10m33s): The stroke had been quick and careless, biting deep. Looking down, Bran glimpsed pale flesh where the wool of his leggings had parted. Then the blood began to flow. He watched the red stain spread, feeling light-headed, curiously apart; there had been no pain, not even a hint of feeling. The big man grunted in surprise.

Figure 7. This figure is best viewed on screen. We present more examples of rich attributes (highlighted) that are extracted from the *Game of Thrones* data set. The location of the description is abbreviated as *chapter number* (Ch), *paragraph number* (P) and the location in the correct video as *episode number* (E) and minutes and seconds. The caption text is directly used from the novel. The various attributes are highlighted in the caption below every frame.