Book2Movie: Aligning Video scenes with Book chapters

Makarand Tapaswi Martin Bäuml Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{makarand.tapaswi, baeuml, rainer.stiefelhagen}@kit.edu

Abstract

Film adaptations of novels often visually display in a few shots what is described in many pages of the source novel. In this paper we present a new problem: to align book chapters with video scenes. Such an alignment facilitates finding differences between the adaptation and the original source, and also acts as a basis for deriving rich descriptions from the novel for the video clips.

We propose an efficient method to compute an alignment between book chapters and video scenes using matching dialogs and character identities as cues. A major consideration is to allow the alignment to be non-sequential. Our suggested shortest path based approach deals with the non-sequential alignments and can be used to determine whether a video scene was part of the original book. We create a new data set involving two popular novel-to-film adaptations with widely varying properties and compare our method against other text-to-video alignment baselines. Using the alignment, we present a qualitative analysis of describing the video through rich narratives obtained from the novel.

1. Introduction

TV series and films are often adapted from novels [4]. In the fantasy genre, popular examples include the *Lord of the Rings* trilogy, the TV series *Game of Thrones* based on the novel series *A Song of Ice and Fire*, the *Chronicles of Narnia* and *Harry Potter* series. There are many other examples from various genres: The *Bourne* series (action), *The Hunger Games* (adventure), the *House of Cards* TV series (drama), and a number of super hero movies (*e.g.* *Batman*, *Spiderman*) many of which are based on comic books.

We believe that such adaptations are a large untapped resource that can be used to simultaneously improve and understand the story semantics for both video and natural language. In recent years there is a rising interesting to automatically generate meaningful captions for images [10, 11, 16, 30] and even user-generated videos [7, 21, 29]. For readers to visualize the story universe, novels often provide rich textual descriptions in the form of attributes or actions of the content that are depicted visually in the video (*e.g.*, specially for characters and their surroundings). Due to the large number of such adaptations of descriptive text into video, such data can prove to be an excellent source for learning joint textual and visual models.

Determining how a specific part of the novel is adapted to the video domain is a very interesting problem in itself [9, 17, 31]. A classical example of this is to pin-point the differences between books and their movie adaptations[1]. Differences are at various levels of detail and can range from appearance, presence or absence of characters, all the way up to modification of the sub-stories. As a start, we address the problem of finding differences at the scene level, more specifically on how chapters and scenes are ordered, and which scenes are not backed by a chapter from the book.

A more fine-grained linkage between novels and films can have direct commercial applications. Linking the two encourages growth in the consumption of both literary and audio-visual material [3].

In this work, we target the problem of aligning video scenes to specific chapters of the novel (Sec. 4). Fig. 1 presents an example of the ground truth alignment between book chapters and parts of the video along with a discussion of some of the challenges in the alignment problem. We emphasize on finding scenes that are not part of the novel and perform our evaluation with this goal in mind (Sec. 5).

There are many applications which emerge from the alignment. Sec. 6 discusses how vivid descriptions in a book can be used as a potential source of weak labels to improve video understanding. Similar to [25], an alignment between video and text enables text-based video retrieval since searching for a part of the story in the video can be translated to searching in the text (novel). In contrast to [25], which used short plot synopses of a few sentences per video, novels provide a much larger pool of descriptions which can potentially match a text-based search. Novels often contain more information than what appears in a short video adaptation. For example, minor characters that appear

[1] There are websites which are dedicated to locating such differences, *e.g.* http://thatwasnotinthebook.com/

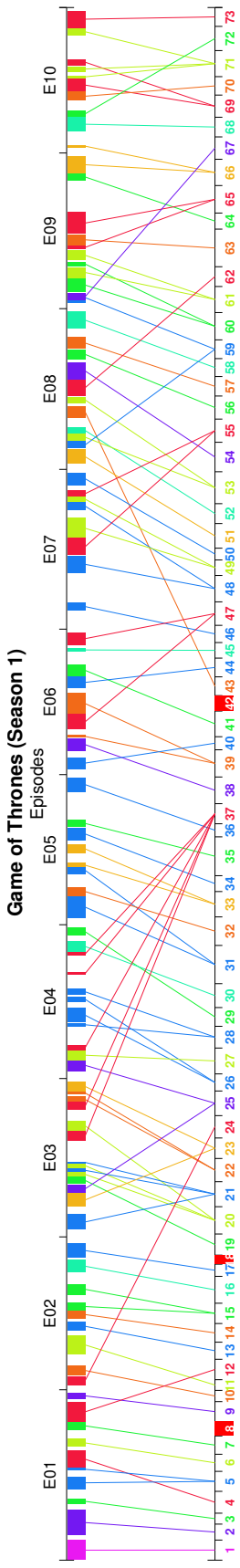


Figure 1. This figure is best viewed in color. We present the ground truth alignment of chapters from the first novel of *A Song of Ice and Fire* to the first season of the TV series *Game of Thrones*. Book chapters are depicted on the lower axis with tick spacing corresponding to the number of words in the chapter. As the novel follows a point-of-view narrative, each chapter is color coded based on the character. Chapters with a red background (see 8, 18, 42) are not represented in the video adaptation. The ten episodes (E01 – E10) of the season 1 corresponding to the first novel are plotted on the top axis. Each bar of color represents the location and duration of time in the video which corresponds to the specific chapter from the novel. A line joining the center of the chapter to the bar indicates the alignment. White spaces between the colored bars indicate that those scenes are not aligned to any part of the novel: almost 30% of all shots do not belong to any chapter. Another challenge is that chapters can be split and presented in multiple scenes (*e.g.*, chapter 37). Note the complexity of the adaptation and how the video does not necessarily follow the book chapters sequentially. We will model these intricacies and use an efficient approach to find a good alignment.

in the video are often un-named, however, can be associated with a name by leveraging the novels. Finally, the alignment when performed on many films and novels together can be used to study the task of adaptation itself and, for example, characterize different screenwriters and directors.

2. Related work

In recent years, videos and accompanying text are an increasingly popular subject of research. We briefly review and discuss related work in the direction of video analysis and understanding using various sources of text.

The automatic analysis of TV series and films has greatly benefited from external sources of textual information. Fan transcripts with subtitles are emerging as the de facto standard in automatic person identification [6, 8, 19, 24] and action recognition [5, 12]. They also see application in video summarization [28] and are used to understand the scene structure of videos [13]. There is also work on aligning videos to transcripts in the absence of subtitles [23].

The last year has seen a rise in joint analysis of text and video data. [22] presents a framework to make TV series data more accessible by providing audio-visual meta data along with three types of textual information – transcripts, episode outline and video summaries. [25] introduces and investigates the use of plot synopses from Wikipedia to perform video retrieval. The alignment between sentences of the plot synopsis and shots of the video is leveraged to search for story events in TV series. A new source of text in the form of descriptive video service used to help visually impaired people watch TV or films is introduced by [20]. In the domain of autonomous driving videos, [14] parses textual queries as a semantic graph and uses bipartite graph matching to bridge the text-video gap, ultimately performing video retrieval. Very recently, an approach to generate textual descriptions for short video clips through the use of Convolutional and Recurrent Neural Networks is presented in [29]. While the video domains may be different, the improved vision and language models have spiked a clear interest in co-analyzing text and video data.

Adapting a novel to a film is an interesting problem in the performing arts literature. [9, 17, 31] present various guidelines on how a novel can be adapted to the screen or theater[2]. Our proposed alignment is a first step towards automating the analysis of such adaptations at a large scale.

Previous works in video analysis almost exclusively deal with textual data which is derived from the video. This often causes the text to be rather sparse owing to limited human involvement in its production. One key difference in analyzing novels – as done in this work – is that the video is derived from the text (or, both are derived from the core

[2] http://www.kvenno.is/englishmovies/adaptation.htm is a nice summary in terms of the differences and freedom that different medium may use while essentially telling the same story.

storyline). This usually means that the available textual description is much more complete which has both advantages (more details) and disadvantages (not every part of the text needs to be present in the video). We think that this ultimately provides a powerful source for improving simultaneous semantic understanding of both text and video.

Closest to our alignment approach are two works which use dynamic programming (DP) to perform text-to-video alignment. Sankar *et al*. [23] align videos to transcripts without subtitles; and Tapaswi *et al*. [25] align sentences from the plot synopses to shots of the video. However, both models make two strong assumptions: (i) sentences and video shots appear sequentially; (ii) every shot is assigned to a sentence. This is often not the case for novel-to-film adaptations (see Fig. 1). We specifically address the problem of non-sequential alignments in this paper. To the best of our knowledge, this is also the first work to automatically analyze novels and films, to align book chapters with video scenes and as a by-product derive rich and unrestricted attributes for the visual content.

3. Data set and preprocessing

We first briefly describe our new data set followed by some necessary pre-processing steps.

3.1. A new Book2Movie data set

As this is the first work in this direction, we create a new data set, comprising of two novel-to-film/series adaptations:

- GOT: Season 1 of the TV series *Game of Thrones* corresponds to the first book of the *A Song of Ice and Fire* series, titled *A Game of Thrones*. Fig. 1 shows our annotated ground truth alignment for this data.
- HP: The first film and book of the *Harry Potter* series – *Harry Potter and the Sorcerer's Stone*. A figure similar to Fig. 1 is included in the supplementary material.

Our choice of data is motivated by multiple reasons. Firstly, we consider both TV series and film adaptations which typically have different filming styles owing to the structure imposed by episodes. There is a large disparity in the sizes of the books (GOT source novel is almost 4 times as big as HP), and this is also reflected in the total runtime of the respective videos (`9h` for GOT vs. `2h30m` for HP).

Secondly, the stories are targeted towards different audiences and thus their language and content differs vastly. The first book of the *Harry Potter* series caters towards children while the same cannot be said about *A Game of Thrones*.

Finally, while both stories are from the fantasy and drama genre, they have their own unique worlds and different writing styles. GOT presents multiple sub-stories that take place in different locations in the world at the same time. This allows for more freedom in the adaptation of the sub-stories creating a complex alignment pattern. On the other hand, HP is very character centric with a large chunk of the story revolving around the main character (Harry). The story and the adaptation are thus relatively sequential.

Some statistics of the data set can be found in Sec. 5.

3.2. Scene detection

We consider shots as the basic unit of video processing and they are used to annotate the ground truth novel-to-film alignment. Due to the large number of shots that films contain (typically more than 2000), we further group shots together and use video *scenes* as basis for the alignment.

We perform scene detection using the dynamic programming method proposed in [26]. The method optimizes the placement of scene boundaries such that shots within a scene are most similar in color and shot threads are not split.

We sacrifice on granularity and induce minor errors owing to wrong scene boundary detection. However, the usage of scenes reduces complexity of the alignment algorithm, and facilitates stronger cues obtained by an average over multiple shots in the scene.

3.3. Film dialog parsing

We extract character dialogs from the video using subtitles which are included in the DVD. We convert subtitles into dialogs based on a simple, yet widely followed set of rules. For example, a two line subtitle whose lines start with "–" are spoken by different characters. Similarly, we also group subtitles appearing consecutively (no time spacing between the subtitles) until the sentence is completed. These subtitle-dialogs are one video cue to perform the alignment.

3.4. Novel dialog parsing

We follow a hierarchical method for processing the entire text of the novel. We first divide the book into chapters and each chapter into paragraphs. For the alignment we restrict ourselves at the level of chapters.

Paragraphs are essentially of two types: (i) with dialog or (ii) with only narration.

Dialogs in paragraphs are indicated by the presence of quotation marks " and ". Each sentence in the dialog is treated separately. These text-based dialogs are used as an alignment cue along with the video dialogs.

The narrative paragraphs usually set the scene, describe the characters, and back stories. They are our major source of attribute labels (see Sec. 6). We process the entire book with part-of-speech tagging using the Stanford CoreNLP [2] toolbox. This provides us with the necessary adjective tags for extracting descriptions.

4. Aligning book chapters to video scenes

We propose a graph-based alignment approach to match chapters of a book to scenes of a video. Fig. 2 presents an

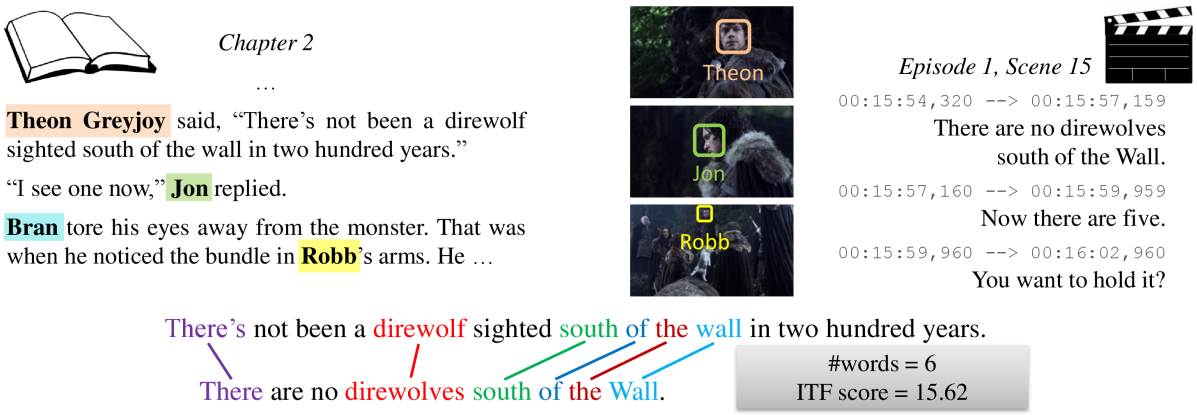


Figure 2. Alignment cues: We present a small section of the GOT novel, chapter 2 (left) and compare it against the first episode of the adaptation (subtitles shown on the right). Names in the novel are highlighted with colors for which we expect to find corresponding face tracks in the video. We also demonstrate the dialog matching process of finding the longest common subsequence of words. The number of words and the score (inverted term frequency) are displayed for one such matching dialog. This figure is best viewed in color.

overview of the cues used to perform the alignment.

For a specific novel-to-film adaptation, let N_S be the number of scenes and N_C the number of chapters. Our goal is to find a chapter c_s^* which corresponds to each scene s

$$c_s^* = \arg\max_{c=\{\varnothing,1\ldots N_C\}} \phi_{(c,s)} \ , \tag{1}$$

subject to certain story progression constraints (discussed later). All scenes which are not part of the original novel are assigned to the $\varnothing$ (null) chapter. $\phi_{(c,s)}$ captures the similarity between the chapter c and scene s.

4.1. Story characters

Characters play a very important role in any story. Similar to the alignment of videos to plot synopses [25], we propose to use name mentions in the text and face tracks in the video to find occurrences of characters.

Characters in videos To obtain a list of names which appear in each scene, we first perform face detection and tracking. Following [27], we align subtitles and transcripts and assign names to “speaking faces”. We consider all named characters as part of our cast list. As feature, we use the recently proposed VF^2 face track descriptor [18] (a Fisher vector encoding of dense SIFT features with video pooling). Using the weakly labeled data obtained from the transcripts, we train one-vs-all linear Support Vector Machine (SVM) classifiers for each character and perform multi-class classification on all tracks. Tracks which have a low confidence towards all character models (negative SVM scores) are classified as “unknown”.

Sec. 5 presents a brief evaluation of face track recognition performance.

Finding name references in the novel Using the list of characters obtained from the transcripts and along with the proper noun part-of-speech tags, finding name mentions in the book is fairly straightforward. However, complex stories – especially with multiple people from the same family – can contain ambiguities. For example, in *Game of Thrones*, we see that Eddard Stark is often referred to as Ned, or addressed with the title Lord Stark. However, as titles can be passed on, his son Robb is also referred to as Lord Stark in the same book. We weight the name mentions based on their types (ordered from highest to lowest): (i) full name; (ii) only first name; (iii) alias or titles; (iv) and only last name. The actual search for names is performed by simple text matching of complete words.

Matching For every scene and chapter, we count the number of occurrences for each character. We stack them as “histograms” of dimension N_P (number of people). For scenes, we count the number of face tracks $\mathbf{Q}_S \in \mathbb{R}^{N_S \times N_P}$, and for chapters we obtain weighted name mentions $\mathbf{Q}_C \in \mathbb{R}^{N_C \times N_P}$. We normalize the occurrence matrices such that the row-sum equals 1, and then compute Euclidean distance between them. Finally, we normalize the distance and obtain our identity based similarity score as

$$\phi^I_{(c,s)} = \sqrt{2} - \|\mathbf{q}_S(s) - \mathbf{q}_C(c)\|^2 \ , \tag{2}$$

where $\mathbf{q}_S(s)$ stands for row s of matrix $\mathbf{Q}_S$. We use this identity-based similarity measure $\phi^I \in \mathbb{R}^{N_C \times N_S}$ between every chapter c and scene s to perform the alignment.

4.2. Dialog matching

Finding an identical dialog in the novel and video adaptation is a strong hint towards the alignment. While matching dialogs between the novel and film sounds easy, often the adaptation changes the presentation style so that very few dialogs are reproduced verbatim. For example, in GOT, we have 12,992 dialogs in the novel and 6,842 in the video.

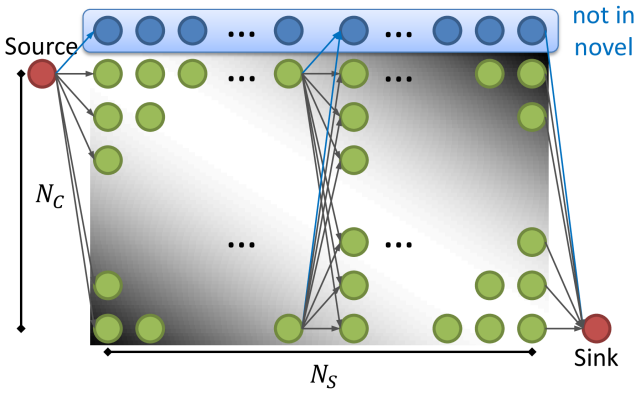


Figure 3. Illustration of the graph used for alignment. The background gradient indicates that we are likely to reach nodes in the lighter areas while nodes in the darker areas are harder to access.

Of these, only 308 pairs are a perfect match with 5 words or more, while our following method finds 1,358 pairs with high confidence.

We denote the set of dialogs in the novel and video by $\mathcal{D}_N$ and $\mathcal{D}_V$ respectively. Let W^N and W^V be the corresponding total number of words. We compute the term frequency [15] for each word w in the video dialogs as

$$\psi^V(w) = \#w / W^V \,, \tag{3}$$

where $\#w$ is the number of occurrences of the word. A similar term frequency $\psi^N(w)$ is computed for the novel.

To quantify the similarity between a pair of dialogs, $n \in \mathcal{D}_N$ from the novel and $v \in \mathcal{D}_V$ from the video, we find the longest common subsequence between them (see Fig. 2) and extract the list of matching words $\mathcal{S}_{n,v}$. However, stop words such as *of, the, for* can produce false spikes in the raw count $|\mathcal{S}_{n,v}|$. We counter this by incorporating inverted term frequencies and compute a similarity score between each dialog pair (n, v) as

$$\phi^D_{(n,v)} = \sum_{w \in \mathcal{S}_{n,v}} -(\log \psi^N(w) + \log \psi^V(w))/2 \,. \tag{4}$$

For the alignment, we trace the dialogs to their respective book chapters and video scenes, and accumulate them

$$\phi^D_{(c,s)} = \sum_{n \in c_D} \sum_{v \in s_D} \phi^D_{(n,v)} \,, \tag{5}$$

where c_D and s_D is the set of dialogs in the chapter and scene respectively. We finally normalize the obtained dialog-based similarity matrix $\phi^D \in \mathbb{R}^{N_C \times N_S}$.

4.3. Alignment as shortest path through a graph

We model the problem of aligning the chapters of a book to scenes of a video (Eq. 1) as finding the shortest path in a sparse directed acyclic graph (DAG). The edge distances of the graph are devised in a way that the shortest path through the graph corresponds to the best matching alignment. As we will see, this model allows us to capture all the non-sequential behavior of the alignment while easily incorporating information from the cues and providing an efficient solution. Fig. 3 illustrates the nodes of the graph along with a glimpse of how the edges are formed.

The main graph (with no frills) consists $N_C \cdot N_S$ nodes ordered as a regular matrix (green nodes in Fig. 3) where rows represent chapters (indexed by c) and columns represent scenes (indexed by s). We create an edge from every node in column s to all nodes in column $s+1$ resulting in a total of $N_C^2 \cdot (N_S - 1)$ edges. These edge transitions allow to assign each scene to any chapter, and the overall alignment is simply the list of nodes visited in the shortest path.

Prior Every story has a natural forward progression which an adaptation (usually) follows, *i.e.* early chapters of a novel are more likely to be presented earlier in the video while chapters towards the end come later. We initialize the edges of our graph with the following distances.

The *local* distance from a node (c, s) to any node in the next column $(c', s+1)$ is modeled as a quadratic distance and deals with transitions which encourage neighboring scenes to be assigned to the same or nearby chapters.

$$d_{(c,s)\to(c',s+1)} = \alpha + \frac{|c' - c|^2}{2 \cdot N_C} \,, \tag{6}$$

where $d_{n \to n'}$ denotes the edge distance from node n to n'. α serves as a non-zero distance offset from which we will later subtract the similarity scores.

Another influence of the prior is a *global* likelihood of being at any node in the graph. To incorporate this, we multiply all incoming edges of node (c, s) by a Gaussian factor

$$g(c, s) = 2 - \exp\left(-\frac{(s - \mu_s)^2}{2 \cdot N_S^2}\right) \,. \tag{7}$$

$\mu_s = \lceil c \cdot N_S / N_C \rceil$ and the $2 - \exp(\cdot)$ ensures the multiplying factor $g \geq 1$. This results in larger distances to go towards nodes in the top-right or bottom-left corners.

Overall, the distance to come to any node (c, s) is influenced by where the edge originates ($d_{n \to (c,s)}$, Eq. 6) and the location of the node itself ($g(c, s)$, Eq. 7).

Unmapped scenes Adaptations provide creative freedom and thus every scene need not always stem from a specific chapter from the novel. For example, in GOT about 30% of the shots do not have a corresponding part in the novel (*c.f.* Fig. 1). To tackle this problem, we add an additional layer of N_S nodes to our graph (top layer of blue nodes in Fig. 3). Inclusion of these nodes allows us to maintain the column-wise edge structure while also providing the additional feature of unmapped scenes. The distance to arrive at this node from any other node is initialized to $\alpha + \mu_g$, where μ_g is the average value of the global influence Eq. 7.

External source/sink Initializing the shortest path within the main graph would force the first scene to belong to the first chapter, and the last scene to the last chapter. However, this need not be true. To prevent this, we include two additional source and sink nodes (red nodes in Fig. 3). We create N_C additional edges each to transit from the source to column $s = 1$ and from column $s = N_S$ to the sink. The distances of these edges are based on the global distance.

Using the alignment cues The shortest path for the current setup of the graph without external influence assigns equal number of scenes to every chapter. Due to its structure, we refer to this as the diagonal prior. We now use the identity (Eq. 2) and dialog matching (Eq. 4) cues to lower the distance of reaching a high scoring node. For example, a node which has multiple matching dialogs and the same set of characters is very likely to depict the same story and be a part of correct scene-to-chapter alignment. Thus reducing the incoming distance to this node encourages the shortest path to go through it.

For every node (c, s) with a non-zero similarity score, we subtract the similarity score from all incoming edges, thus encouraging the shortest path to go *through* this node.

$$d_{n' \to (c,s)} = d_{n' \to (c,s)} - \sum_M w_M \phi^M_{(c,s)} \,, \tag{8}$$

where $\phi^M_{(\cdot)}$ is the similarity score of modality M with weight $w_M > 0$ and $n' = (\cdot, s-1)$ is the list of nodes from the previous scene with an edge to (c, s). The sum of weights across all modalities is constrained as $\sum_M w_M = 1$. In our case, the modalities comprise of dialog matching and character identities, however, such an approach allows to easily extend the model with more cues.

When the similarity score between scene s and all chapters is low, it indicates that the scene might not be a part of the book. We thus modify the incoming distance to the node $(\varnothing, s)$ as follows:

$$d_{n' \to (\varnothing,s)} = d_{n' \to (\varnothing,s)} - \sum_M w_M \max\left(0, 1 - \sum_c \phi^M_{(c,s)}\right) . \tag{9}$$

This encourages the shortest path to assign the scene s to the null chapter class $c^*_s = \varnothing$.

Implementation details We use Dijkstra's algorithm to solve the shortest path problem. For GOT, our graph has about 27k nodes and 2M edges. Finding the shortest path is very efficient and takes less than 1 second. Our initialization distance parameter α is set to 1. The weights w_M are not very crucial and result in good alignment performance for a large range of values.

5. Evaluation

We now present evaluation of our proposed alignment approach on the data set [1] described in Sec. 3.1.

VIDEO		duration	#scenes	#shots (#nobook)
	GOT	8h 58m	369	9286 (2708)
	HP	2h 32m	138	2548 (56)
BOOK		#chapters	#words	#adj, #verb
	GOT	73	293k	17k, 59k
	HP	17	78k	4k, 17k
Face-ID		#characters	#tracks (unknown)	id accuracy
	GOT	95	11094 (2174)	67.6
	HP	46	3777 (843)	72.3

Table 1. An overview of our data set. The table is divided into three sections related to information about the *video*, the *book* and the *face identification* scheme.

Data set statistics Table 1 presents some statistics of the two novel-to-video adaptations. The GOT video adaptation and book is roughly four times larger than that of HP. A major difference between the two adaptations is the fraction of shots in the video which are not part of a book chapter. For GOT the number of shots not in the book (#nobook) are much higher at 29.2% than as compared to only 2.2% for HP. Both novels are a large resource for adjectives (#adj) and verbs (#verb).

Face ID performance Table 1 (*Face-ID*) shows the overall face identification performance in terms of track-level accuracy (fraction of correctly labeled face tracks).

The face tracks involve a large number of people and many unknown characters, and our track accuracy of around 70% is on par with state-of-the-art performance on complex video data [18]. Additionally, GOT and HP are a new addition to the existing data sets on person identification and we will make the tracks and labels publicly available [1].

5.1. Ground truth and evaluation criterion

The ground truth alignment between book chapters and videos is performed at the *shot* level. This provides a fine-grained alignment independent of the specific scene detection algorithm. The scene detection only helps simplify the alignment problem by reducing the complexity of the graph.

We consider two metrics. First, we measure the alignment accuracy (*acc*) as the fraction of shots that are assigned to the correct chapter. This includes shot assignments to $\varnothing$. We also emphasize on finding shots that are not part of the book (particularly for GOT). Finding these shots is treated like a detection problem, and we use precision (*nb-pr*) and recall (*nb-rc*) measures.

5.2. Book-to-video alignment performance

We present the alignment performance of our approach in Table 2, inspect the importance of dialog and identity cues, and compare the method against multiple baselines.

As discussed in Sec. 3.2, for the purpose of alignment, we perform scene detection and group shots into scenes.

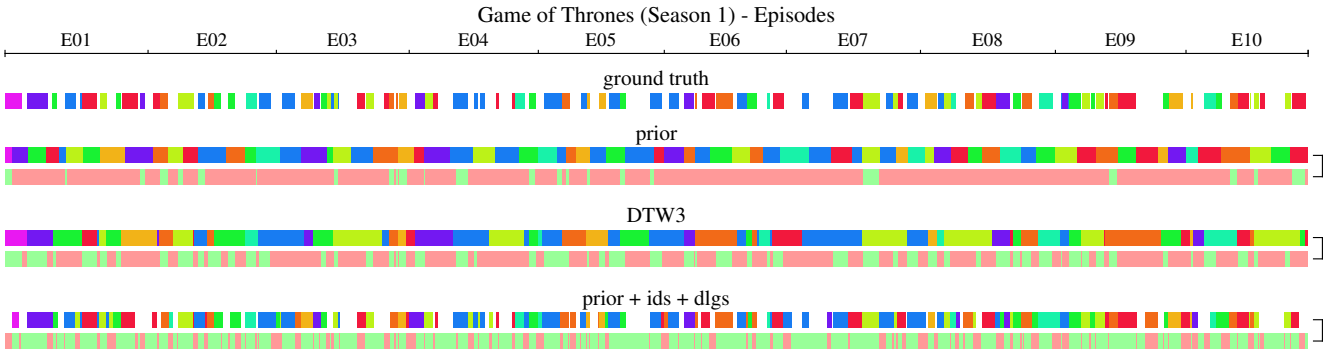


Figure 4. This figure is best viewed in color. We visualize the alignment as obtained from various methods. The ground truth alignment (row 1) is the same as presented in Fig. 1. Chapters are indicated by colors, and empty areas (white spaces) indicate that those shot are not part of the novel. We present 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).

method	GOT			HP		
	acc	nb-pr	nb-rc	acc	nb-pr	nb-rc
scenes upper	95.1	97.9	86.4	96.7	40.0	7.1
prior	12.4	–	–	19.0	–	–
prior + ids	55.3	52.8	48.7	80.4	0.0	0.0
prior + dlgs	73.1	55.8	**74.2**	86.2	20.0	3.6
ids + dlgs	66.5	71.7	20.9	77.4	0.0	0.0
prior + ids + dlgs	**75.7**	**70.5**	53.4	**89.9**	0.0	0.0
MAX [25]	54.9	–	–	73.3	–	–
MAX [25]+∅	**60.7**	68.0	37.7	73.0	0.0	0.0
DTW3 [25]	44.7	–	–	**94.8**	–	–

Table 2. Alignment performance in terms of overall accuracy of shots being assigned book chapters (acc). The no-book precision (nb-pr) and no-book recall (nb-rc) are for shots which are not part of the book. See Sec. 5.2 for a detailed discussion.

However, errors in the scene boundary detection can lead to small alignment errors measured at the shot level. Method *scenes upper* is the best possible alignment performance given the scene boundaries and we observe that we do not lose much in terms of overall accuracy.

Shortest-path based As a baseline (*prior*), we evaluate the alignment obtained from our graph after initializing edge distances with local and global priors. The alignment is an equal distribution of scenes among chapters. For both GOT and HP we observe bad performance suggesting the difficulty of the task. The prior distances modified with the dialog matching cue (*prior + dlgs*) outperform prior with character identities (*prior + ids*) quite significantly. An explanation of this is (i) inaccurate face track recognition; and (ii) similar names appearing in the text (*e.g.* Jon Arryn, Jon Umber and Jon Snow are three characters in GOT). However the fusion of the two cues *prior + ids + dlgs* performs best. Note that we are quite successful in finding shots that are not part of the book. If we disable the global prior which provides structure to the distances (*ids + dlgs*), we obtain comparable performance indicating that the cues are responsible for the alignment accuracy. In general, our ability to predict whether a shot belongs to the novel or not is good for GOT. However, for HP as the no-book shots are very few (2.2%), they are difficult to discern.

The accuracy is robust against a large range of weights. For example, for GOT, sweeping $w_{id} \in [0.01, 0.5]$ results in accuracy ranging from 71.5% to 69.9% peaking in between.

Baselines The MAX baseline [25] uses both cues, however treats scenes independent of one another. The method selects the highest scoring chapter for each scene in the similarity matrix $\phi_{(c,s)}$, and performs worse as compared to our method. When augmented with the null class (MAX + ∅) by scoring it similar to Eq. 9 we see an improvement for GOT, while HP performs slightly worse. Note that our proposed method *prior+ids+dlgs* is far better than *MAX*.

DTW3 [25] (similar to DP [23]) forces a scene s to be assigned to the same or subsequent chapter as scene $s-1$. This restriction allows the method to perform well in the case of simple structures (HP), but fails completely in the case of complex alignments (GOT).

Qualitative analysis Fig. 4 presents the alignment obtained by different methods on GOT. Notice how our method is quite successful at predicting scenes which are not part of the book (the white spaces in the first sub-row). Both prior and DTW3 have large chunks of erroneous scenes (second sub-row in red) due to long-range influences of the alignment. A drawback of our method is that it needs sufficient evidence to assign a scene to a chapter. This is specially seen in HP (analysis in supplementary material).

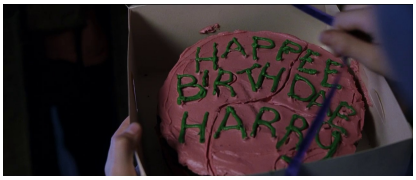
6. Mining rich descriptions

A novel portrays the visual world through rich descriptions. While there are many levels of such a presentation

(a) (Ch11, P47, E02 12m23s): Arya was in her room, packing a polished ironwood chest that was bigger than she was. Nymeria was helping. Arya would only have to point, and the wolf would bound across the room, snatch up some wisp of silk in her jaws, and fetch it back.

(b) (Ch60, P66, E09 7m55s) Lord Walder was ninety, a wizened pink weasel with a bald spotted head, too gouty to stand unassisted. His newest wife, a pale frail girl of sixteen years, walked beside his litter when they carried him in. She was the eighth Lady Frey.

(c) (Ch12, P33, E01 51m21s) One egg was a deep green, with burnished bronze flecks that came and went depending on how Dany turned it. Another was pale cream streaked with gold. The last was black, as black as a midnight sea, yet alive with scarlet ripples and swirls.



(d) (Ch4, P23, M 14m23s) From an inside pocket of his black overcoat he pulled a slightly squashed box. Harry opened it with trembling fingers. Inside was a large, sticky chocolate cake with Happy Birthday Harry written on it in green icing.

(e) (Ch9, P194, M 1h02m45s) They were looking straight into the eyes of a monstrous dog, a dog that filled the whole space between ceiling and floor. It had three heads. Three pairs of rolling, mad eyes; three noses, twitching and quivering in their direction.

(f) (Ch10, P78-79, M 1h06m59s) 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 5. Example of mined attributes ranging from events, scenes, and person and object descriptions. The first row corresponds to GOT, while the second shows examples from HP. The location of the description in the novel is abbreviated as *chapter number* (Ch), *paragraph number* (P) and the location in the video as episode E*XX* or movie M in minutes and seconds. The caption text is verbatim from the novel. The various attributes are highlighted in the caption below every frame. This figure is best viewed on screen.

we highlight and focus on a few key ones: (i) narrative text describing the scene or location; (ii) detailed character descriptions including face-related features, clothing, and also portrayal of their mental state; (iii) character interactions, their emotions and usage of various communication verbs to express a sentiment.

As an application of the scene-to-chapter alignment, we present a sample of attributes that are easily obtained. Fig. 5 presents qualitative examples from our data set. The figure captions are verbatim paragraphs from the novel. For simplicity, we highlight the relevant attributes of the text which correspond to the part of the image. Many more such examples can be found in the supplementary material.

We extract these examples pseudo-automatically from a given scene-to-chapter alignment. Small regions (± 3 paragraphs, ± 5 shots) surrounding matched dialogs are used to find appropriate examples. Note that the video clip, *i.e*. the region of shots around the matching dialog are part of the same story, however we select one good frame for presentation purposes. Although this currently requires some manual intervention, generating pairs of images and rich descriptions is far easier than going through the entire book. Automating this task is an interesting area for future work.

7. Conclusion

With a rising interest in transforming novels to films or TV series, we present for the first time an automatic way to analyze novel-to-film adaptations. We create and annotate a data set involving one film and one season of a TV series. Our goal is to associate video scenes with book chapters and is achieved by modeling the task as a shortest path graph problem. Such a model allows for a non-sequential alignment and is able to handle scenes that do not correspond to parts of the book. We use information cues such as matching dialogs and character identities to bridge the gap between the text and video domains. A comparison against state-of-the-art plot synopsis alignment algorithms shows that we perform much better specially in case of complicated alignment structures. As an application of the alignment between book chapters and video scenes, we also present qualitative examples of extracting rich scene, character, and object descriptions from the novel.

Acknowledgments This work was supported by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) under contract no. STI-598/2-1. The views expressed herein are the authors' responsibility and do not necessarily reflect those of DFG.

References

[1] Book2Movie data download containing novel-to-film ground truth alignments, and face tracks + labels. http://cvhci.anthropomatik.kit.edu/projects/mma. 6

[2] Stanford CoreNLP. http://nlp.stanford.edu/software/. Retrieved 2014-11-14. 3

[3] 'Deathly Hallows' film breathes life into Harry Potter book sales. http://www.nielsen.com, Nov. 2010. Retrieved 2014-11-14. 1

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

[5] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding Actors and Actions in Movies. In *ICCV*, 2013. 2

[6] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *CVPR*, 2009. 2

[7] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *CVPR*, 2013. 1

[8] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" - Automatic Naming of Characters in TV Video. In *BMVC*, 2006. 2

[9] R. Giddings, K. Selby, and C. Wensley. *Screening the novel: The theory and practice of literary dramatization*. Macmillan, 1990. 1, 2

[10] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. 1

[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. 1

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *CVPR*, 2008. 2

[13] C. Liang, C. Xu, J. Cheng, and H. Lu. TVParser : An Automatic TV Video Parsing Method. In *CVPR*, 2011. 2

[14] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *CVPR*, 2014. 2

[15] C. D. Manning, P. Raghavan, and H. Schutze. Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*. Cambridge University Press, 2008. 5

[16] 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. 1

[17] B. McFarlane. *Novel to Film: an Introduction to the Theory of Adaptation*. Clarendon press, Oxford, 1996. 1, 2

[18] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A Compact and Discriminative Face Track Descriptor. In *CVPR*, 2014. 4, 6

[19] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *ECCV*, 2014. 2

[20] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A Dataset for Movie Description. In *CVPR*, 2015. 2

[21] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating Video Content to Natural Language Descriptions. In *ICCV*, 2013. 1

[22] A. Roy, C. Guinaudeau, H. Bredin, and C. Barras. TVD: A Reproducible and Multiply Aligned TV Series Dataset. In *Language Resources and Evaluation Conference*, 2014. 2

[23] P. Sankar, C. V. Jawahar, and A. Zisserman. Subtitle-free Movie to Script Alignment. In *BMVC*, 2009. 2, 3, 7

[24] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – Learning person specific classifiers from video. In *CVPR*, 2009. 2

[25] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based Video Retrieval in TV series using Plot Synopses. In *ACM Intl. Conf. on Multimedia Retrieval (ICMR)*, 2014. 1, 2, 3, 4, 7

[26] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. StoryGraphs: Visualizing Character Interactions as a Timeline. In *CVPR*, 2014. 3

[27] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Improved Weak Labels using Contextual Cues for Person Identification in Videos. In *IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG)*, 2015. 4

[28] T. Tsoneva, M. Barbieri, and H. Weda. Automated summarization of narrative video on a semantic level. In *Intl. Conf. on Semantic Computing*, 2007. 2

[29] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *North American Chapter of the Association for Computational Linguistics Human Language Technologies*, 2015. 1, 2

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015. 1

[31] G. Wagner. *The novel and the cinema*. Fairleigh Dickinson University Press, 1975. 1, 2