Noname manuscript No.
(will be inserted by the editor)

Aligning Plot Synopses to Videos for Story-based Retrieval

Makarand Tapaswi · Martin Bäuml · Rainer Stiefelhagen

Received: date / Accepted: date

Abstract We propose a method to facilitate search through the storyline of TV series episodes. To this end, we use human written, crowdsourced descriptions – *plot synopses* – of the story conveyed in the video. We obtain such synopses from websites such as Wikipedia and propose various methods to align each sentence of the plot to shots in the video. Thus, the semantic story-based video retrieval problem is transformed into a much simpler text-based search. Finally, we return the set of shots aligned to the sentences as the video snippet corresponding to the query.

The alignment is performed by first computing a similarity score between every shot and sentence through cues such as character identities and keyword matches between plot synopses and subtitles. We then formulate the alignment as an optimization problem and solve it efficiently using dynamic programming. We evaluate our methods on the fifth season of a TV series *Buffy the Vampire Slayer* and show encouraging results for both the alignment and the retrieval of story events.

Keywords Story-based retrieval · Text-Video alignment · Plot synopsis · TV series

Makarand Tapaswi
Computer Vision for Human Computer Interaction Lab
Karlsruhe Institute of Technology, Germany
Tel.: +49-721-608-44735
E-mail: makarand.tapaswi@kit.edu

Martin Bäuml
E-mail: baeuml@kit.edu

Rainer Stiefelhagen
E-mail: rainer.stiefelhagen@kit.edu

1 Introduction

Searching for people, actions, events, concepts, and stories in large-scale video content is a very challenging problem. Many popular tasks have been proposed under the TRECVid [38] evaluation campaign. Two related ones, multimedia event detection (MED) and semantic indexing (SIN) focus on concept detection [40] and work primarily in the domain of unstructured user-generated videos. Concepts in SIN correspond to events (*e.g. Election Campaign*), scenes (*e.g. Snow*), person descriptions (*e.g. Male Anchor*), objects (*e.g. Sofa*), and simple actions (*e.g. Throw Ball*). On the other hand, events in MED attempt to cover an informative spectrum of videos with mid-level descriptions (*e.g. Attempting a bike trick*, *Repairing an appliance*). In contrast to using low-level features such as color and texture for retrieval, both MED and SIN attribute a higher semantic meaning to the videos. However, they are quite far from automatically understanding or interpreting the video content at the level of storytelling.

A large amount of work exists in the domain of broadcast news, debates, and talk shows. The major focus is on indexing the videos for further analysis, primarily for retrieval applications. Some of the topics here are speaker identification [32], analysis and fusion of face and speech cues to perform audio-visual person identification [9], detection of stories [31], and segmentation of narratives [23] in news.

Our focus, however, is primarily on professionally edited videos produced to convey a story such as TV series and movies. The recent advances in this domain are mainly in popular computer vision topics such as action recognition [21], person identification [6,10,37], pose estimation [5], and human interaction analysis [30]. Former challenges such as shot and scene change detec-

tion [33], and video copy detection [22] are giving way to new tasks which promise a higher level of abstraction and move towards better understanding of videos. Some such examples include MediaEval's violent scenes detection challenge [12], visualization and grouping scenes which belong to the same story threads [13], and visualization of TV episodes as a chart of character interactions [44].

Nevertheless, there is an interpretation gap between all the automatically generated metadata (person identities, scenes, actions, etc.) and the actual storyline of these videos. Even with all the metadata, searching for a specific plot within the story such as "Darth Vader tries to convince his son Luke to join him" (source – *Star Wars - The Empire Strikes Back*) or "Golem succeeds in snatching the ring from Frodo" (source – *Lord of the Rings - The Return of the King*) is a challenging problem. In this paper, we propose to use crowdsourcing in the form of human written descriptions (plot synopses) on Wikipedia or other fan sites to address the problem of story-based video retrieval.

Prior to the retrieval, we first align sentences within the plot to the shots of the video. To guide the alignment, we propose to use entities that appear both in the visual (episode or movie) and textual (plot) depictions of the story. Such elements include characters within the story, the locations, actions, objects, and events. Characters form a major part in shaping any story as the story essentially revolves around character interactions.

The alignment of textual descriptions to the corresponding shots in a video opens up novel ways to approach some existing applications in the field of video analysis that are otherwise difficult to achieve while relying on video content only. Some applications could be *semantic video summarization* and *automatic description of videos*. A text summarization approach (automatic or manual) applied on the plot synopsis can be used to first select important sentences of the storyline. The alignment then allows to select shots from the complete video, on which standard video summarization techniques [25] can be subsequently applied. In the domain of video understanding, a very important aspect is the ability to automatically generate high-level descriptions such as plot synopses. The alignment between existing plots and videos can be seen as a first step in this direction and used to model the relation between videos and their textual description. Note that the above applications are out of scope of the current work which focuses on the problem of video retrieval.

This paper is an extension of our previous work [43] and presents more insights and discussions on the results. The main contributions of this paper are an approach to perform alignment between human written descriptions (plot synopses) and shots in the video, and demonstration of the obtained alignment on the task of story-based video retrieval.

The paper is presented as follows. First, we discuss related work in Sec. 2, followed by a short analysis of the pre-processing steps required for both the modalities: text and video (Sec. 3.1). We discuss extraction of character identities in Sec. 3.2, and the use of subtitles in Sec. 3.3 as cues to guide the alignment process. Various techniques are proposed to perform the alignment in Sec. 4. Sec. 5 discusses the approach to use the alignment to perform retrieval. We evaluate the performance of alignment in Sec. 6.3 and analyze retrieval results in Sec. 6.4. Finally we present our conclusions and directions for future work in Sec. 7.

2 Related Work

We present an overview of the related work in three broad areas (i) crowdsourcing in video retrieval and summarization, (ii) alignment of videos to various forms of textual descriptions, and a short overview on (iii) automatic generation of image and video descriptions.

2.1 Video Retrieval, Summarization and the role of Crowdsourcing

Over the years TRECVid [38] has been the primary evaluation campaign for video retrieval through tasks such as MED and SIN. The major shift in the video retrieval perception can be attributed to the jump from low-level content features to concept-based video retrieval [39].

Video summarization too has moved from low-level visual features towards semantic content, specifically targeting character identities in the stories. Sang and Xu [35] perform summarization using the structure – shots, scenes, substories – of movies and TV episodes along with the influence of characters. Tsoneva *et al.* [45] use textual information like subtitles and transcripts to help improve summarization by spotting main characters names and their presence in the storyline. Towards this goal of using semantic content for summarization, we believe that plot synopses can have an important role to play. The plots serve as a high-level interpretation of the story, and as discussed in the introduction, can help improve the information content of the generated video summary.

Crowdsourcing specially since the introduction of Amazon Mechanical Turk is gaining popularity in many image/video tasks. For example, Freiburg *et al.* [16]

present a system to easily navigate within concert videos by augmenting them with concert related concepts which indicates the content shown in the shot – *Singer* or *Keyboard*. The automatic concept detection is enhanced by a user feedback system. In the image search domain crowdsourcing in the form of Wikipedia articles is used to learn a joint latent space of image and text pairs for topic models [46]. In the domain of video summarization, crowdsourcing has been used in a novel way to automate the difficult and time-consuming task of evaluating various summarization outputs [20].

2.2 Text to Video Alignment

Using information from text sources such as subtitles and transcripts is a relatively common feature for analysis of TV series or movies. The alignment between transcripts and subtitles has historically provided the means for mining weak labels for person identification tasks [6, 14, 37]. Action recognition has also used transcripts [21] which not only contain names and dialogs, but also include information describing low-level actions of the characters, *e.g.* "He *sits* in the car".

Alignment of transcripts to videos when no subtitles are available is an interesting problem. Working in this area, Sankar *et al.* [36] rely on visual features such as faces (or characters), locations, and the output of an automatic speech recognition system to perform the alignment.

Even in different domains such as sports videos, Xu *et al.* [47] have used webcast text (a sort of transcript) for event detection. However the alignment is relatively easy since the webcast contains timestamps, and most sports videos display the game time next to the current score.

Similarly, in the domain of car-centric videos obtained from driving in the city, a very recent work, Lin *et al.* [27] proposes to use natural language queries to perform semantic search. They first obtain and parse descriptions for the videos into a semantic graph and align the text to video using bipartite graph matching. Object appearance, motion, and spatial relations are captured in their descriptions and constitute the cues for the matching.

Back to TV series, an interesting application of the alignment of videos to transcripts is to generate new videos [26]. Through the use of transcripts, they first index a large video database with characters, place, and timing information. Given a new script they use this metadata followed by post-production to automatically generate new videos.

In general, note that the alignment of transcripts to videos via subtitles is much easier than plot synopses as transcripts always contain dialog information, while the plot describes the story in a concise dialog free manner.

Here is an example excerpt from a transcript[1]

> GILES: Thank you, Willow. Obstinate bloody machine simply refused to work for me. (Walks off)
> WILLOW: Just call me the computer whisperer. (Stands up, putting something in the scanner) Let's get scannin'. I want to see this puppy go.
> Giles puts a pile of old books on her outstretched arms.
> GILES: Start with those.

and the corresponding plot synopsis[2] text.

> Giles has Willow start scanning books into a computer so that they can be resources for the gang to use.

Note how a single sentence from the plot synopsis summarizes the dialog involving the two characters.

2.3 Automatic Image and Video Description

While humanlike description of *any* video is a very challenging problem, there is some work on understanding specific domains of video.

In the domain of surveillance footage (street and car scenes) [29] provides an overview of the effort to "understand what is going on". A Fuzzy Metric Temporal Logic is used to represent both schematic and instantiated knowledge along with its development over time. Together with Situation Graph Trees a natural language description is generated.

In recent years, Gupta *et al.* [17] use action recognition and model the events in sports videos (baseball) by an AND-OR graph. Tan *et al.* [41] use audio-visual concept classifiers and rule-based methods to generate descriptions for a few set of handpicked concepts. Extending this, Habibian and Snoek [18] demonstrate conversion of videos to sentences and vice-versa through a large number of concepts which bridge the gap between text and video.

In this paper, we employ character identities as our "concepts" and use the structure in TV series to align shots to sentences. We also use subtitles as a set of complementary cues to compensate the lack of fully developed vision systems for scene or action recognition in such data. The retrieval is performed by first matching the query to the text and retrieving the corresponding shots from the video.

[1] buffyworld.com/buffy/transcripts/079_tran.html
[2] en.wikipedia.org/wiki/Buffy_vs._Dracula#Plot

3 Text-Video Alignment Cues

The alignment of sentences in plot synopses to shots in the video forms the basis of our retrieval of story events. A graphical overview of the alignment problem is presented in Figure 1.

The primary goal of the alignment is to determine for each sentence s_i of the plot synopsis, a set of Q shots $T_i = \{t_{i1}, \ldots, t_{iQ}\}$ that correspond to the part of the story described in the text. To facilitate the alignment, we formulate a similarity function between every sentence to shot $f(s_i, t_j)$ based on cues arising from person identities and matching keywords in the subtitles.

3.1 Atomic units

We consider shots in the video and sentences from the plot synopsis as the smallest units to perform the alignment.

Plot Synopsis Most movies and TV series episodes have Wikipedia articles which contain a section describing the story of the video in a concise fashion – the plot synopsis. Other sources include fan wiki-sites created for specific TV series such as `http://bigbangtheory.wikia.com`. As the first step in the processing chain, we perform part-of-speech tagging on the sentences of the plot synopsis using the Stanford CoreNLP [2] software suite. A list of characters in the episode, obtained from sources such as IMDb or Wikipedia, is compared against the proper nouns (NNP) to determine the occurrence of characters in the text. Such a sentence augmented with the above information forms the smallest unit for our alignment.

Video A video shot is the counterpart of a sentence and is the atomic unit for the alignment. We perform shot detection using a normalized version of the Displaced Frame Difference (DFD) [48]

$$DFD(t) = \|F(x, y, t) - F((x, y) + D(x, y), t - 1)\|. \quad (1)$$

$D(x, y)$ is the optical flow between frames $F(x, y, t)$ and $F(x, y, t - 1)$ and the DFD computes the motion-compensated difference between them. To detect shot boundaries, we need to find peaks in the DFD scores. We filter the DFD via a top-hat morphological operation and threshold the resulting vector to determine shot boundaries.

We also extract subtitles from the video through OCR [3] and collect transcripts from fan websites. As a minimal requirement, the transcripts should contain *who speaks what* which is used to perform unsupervised person identification as described in the following.

3.2 Character Identification

Character interactions form the backbone of any storytelling. For example, this is used in tasks such as video summarization [35] where person identities are used to influence importance of shots.

We show via experiments that character identities are influential in the alignment of plot synopsis sentences to video shots (*cf*. Sec. 6.3.1). When a character is mentioned in the text, it is highly likely that he/she appears on screen in that storyline. In the structure of any sentence as "subject – verb – object", we observe that the subject, and often even the object refers to characters in the storyline.

Identity extraction from text. To resolve pronouns and other character references (*e.g.* sister, father, etc.), we perform coreference resolution [24] and cluster the nouns attaching them with a name. This is augmented by a simple, yet surprisingly effective technique of looking back for the antecedent that agrees in gender [11]. For example, in this sample from a plot synopsis

> Buffy awakens to find Dracula in her bedroom. She is helpless against his powers and unable to stop him.

we see that *She* and *her* refers to *Buffy* and *his* to *Dracula*.

Identity extraction from video. Person identification in structured videos such as TV series is a popular problem in computer vision [6, 11, 14].

Similar to [6, 14], we perform automatic identification by obtaining weak supervision from subtitles and transcripts. As both subtitles (dialogs with timestamps) and transcripts (dialogs with names) share dialogs, we match the words within dialogs to align subtitles and transcripts and obtain *who* speaks *when* (name and time). We then tag speaking face tracks with the corresponding name. This provides us with labels for roughly 20% of all tracks at a precision of 90%. Instead of nearest-neighbour matching like in [14], we use tracks as weakly labeled training data and train second order polynomial kernel SVM classifiers for each character in a 1-vs-all fashion [42]. We then score tracks against all SVM models and label the track with the character whose SVM scores highest.

Sec. 6.2 evaluates the quality of the identity extraction methods.

Identity similarity function. Different characters appear for different amounts of time in a video. Primary characters are often given large amounts of screen time, appear throughout the video and are referenced frequently

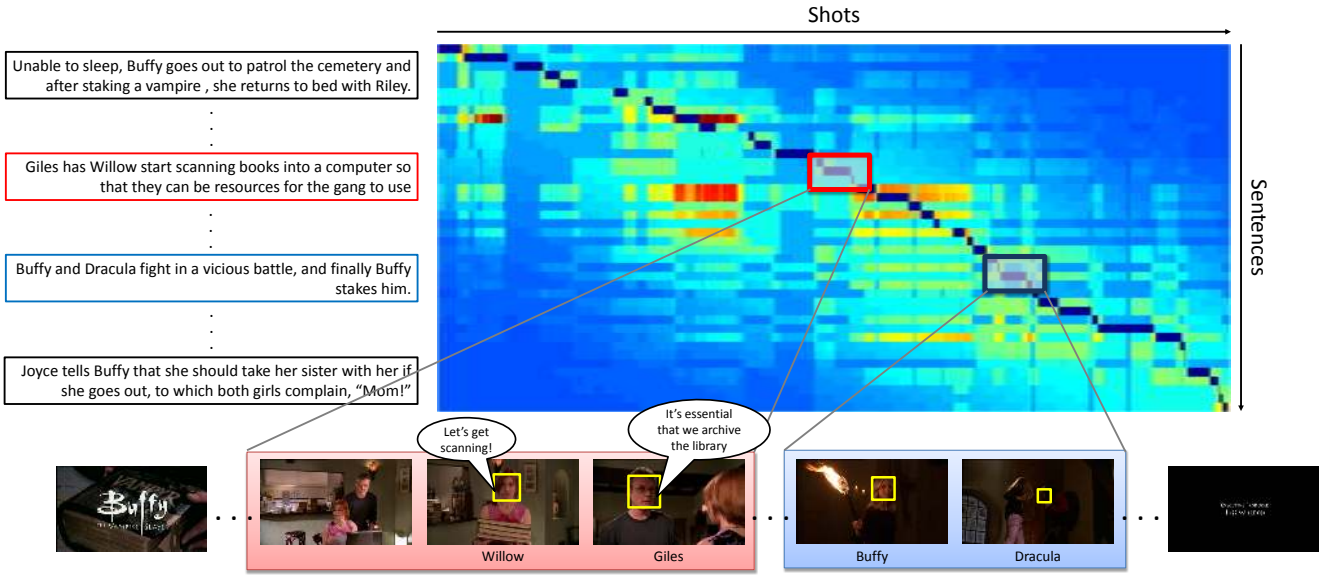

Fig. 1 Sentences from the plot synopsis are aligned to shots from the video. We visualize the similarity matrix $f(s_i, t_j)$ overlayed with the ground truth alignment in dark blue. Each row of the matrix corresponds to a sentence, and each column to a shot. We also present examples of sentences aligned to shots corresponding to same color codes (red and blue). The names of characters in the shot and the dialogs (subtitles) help guide the alignment. This figure is best viewed in color.

in the text. This makes them a bad source for pinpointing shots to sentences. In contrast, guest appearances tend to be less frequent, are given short screen time and are barely mentioned in the text. Thus, when they do actually appear, we obtain a strong hint to align the shot with the corresponding sentence.

We model the importance of each character c^* as

$$\mathcal{I}(c^*) = \frac{\log(\max_{c \in \mathcal{C}} n_{FT}(c))}{\log(n_{FT}(c^*) + 1)}, \tag{2}$$

where $n_{FT}(c)$ is the number of tracks assigned to c and $\mathcal{C}$ is the set of all characters. The importance can be seen as a form of *Inverse Document Frequency* [19].

In many cases we observe that not all characters involved in the storyline are visible in the same shot. Standard editing practices used to create TV series or movies tend to focus the camera on the *speaker* while looking *over the shoulder* of the other nonspeaking character. We observe that spreading the appearance of characters in a small neighborhood of shots is often beneficial to improve alignment. Thus, if character c appears in shot j, we spread his/her influence to a few neighboring shots $j-r, \ldots, j, \ldots, j+r$. We empirically choose $r = 4$.

Finally, the similarity function to match identities between a sentence s_i and shot t_j is given by

$$f_{id}(s_i, t_j) = \sum_{k=j-r}^{j+r} \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} \mathbb{1}\{c = d\} \cdot \mathcal{I}(c), \tag{3}$$

where $\mathcal{C}$ is the set of characters seen in the r neighborhood of shot j and $\mathcal{D}$ is the list of names obtained from sentence i. The term $\mathcal{I}(c)$ is added iff $c = d$.

3.3 Subtitles

In addition to character identities, we use *subtitles* as a cue to align shots of the video to sentences. Note that unlike subtitle-transcript alignment, plot synopses do *not* contain dialogs and describe the summary of the story in the video making the alignment problem much more complicated. Nevertheless, we find a few matches in keywords such as names, places or object references which allow to guide the alignment. While most of the above keywords can also be found using vision tasks (scene recognition, object detection, action recognition, *etc.*), their detection is challenging and tends to introduce additional errors.

We work with shots as atomic units of our video and first assign subtitles to shots via their timestamps. Subtitles which occur at shot boundaries are assigned to the shot which has a majority portion of the subtitle. Prior to the alignment, we normalize the two forms of text by performing stop word removal [8] which induces spurious matches. We compute a similarity function (similar to f_{id} used for character identities) between every sentence s_i from the plot synopses to shot t_j by counting the number of matches between words v in sentence s_i

with w in the subtitles that are assigned to shot t_j

$$f_{subtt}(s_i, t_j) = \sum_{v \in s_i} \sum_{w \in subtt \in t_j} \mathbb{1}\{v = w\} \quad . \tag{4}$$

The resulting similarity matrix f_{subtt} is quite sparse (roughly 7% non-zero entries in one example episode).

3.4 Cue Fusion

The matched keywords between subtitles and plot synopsis typically consist of not only names, but also actions, places, and objects. Along with the character identities, they provide complementary information. We use a simple weighted linear combination of the two similarity functions

$$f_{fus}(s_i, t_j) = f_{id}(s_i, t_j) + \alpha \cdot f_{subtt}(s_i, t_j) \tag{5}$$

where α is chosen to trade-off between the informativeness of subtitles and character identities. We show in our experiments that the fusion demonstrates best performance.

4 Alignment

Given a similarity score between every shot to every sentence, we now turn towards finding a good alignment between shots and sentences. As a general form, we propose the task as an optimization problem over all possible shot-sentence assignments $\mathcal{M} \in (\mathcal{S} \times \mathcal{T})$ where $\mathcal{S}$ is the set of sentences and $\mathcal{T}$ is the set of all possible combinations of shots. In particular, we are interested in finding an optimal assignment $\mathcal{M}^*$ that maximizes the joint similarity $\mathcal{J}(\mathcal{M})$ between shot to sentence alignment.

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}}\, \mathcal{J}(\mathcal{M}) \tag{6}$$

$$= \underset{\mathcal{M}}{\operatorname{argmax}} \left[g(\mathcal{M}) \cdot \sum_{(\mathcal{S},\mathcal{T}) \in \mathcal{M}} f(\mathcal{S}, \mathcal{T}) P(\mathcal{S}, \mathcal{T}) \right] \quad , \tag{7}$$

where $g(\cdot)$ is a generic function that imposes global assignment constraints, for example to not assign every sentence to all shots. Otherwise, if the similarity functions are strictly non-negative (like the ones we use from the Sec. 3) the joint maximum is reached by assigning all shots to every sentence – a trivial solution.

$P(\cdot, \cdot)$ acts as a prior operating on the similarity functions and prevents unexpected behavior such as assignment of the first sentence to last 10 shots of the video, and vice versa.

In practice as the plot is usually a summarized version of the video, the number of sentences N_S is much smaller than the number of shots N_T. Thus, we define

$$g(\mathcal{M}) = \begin{cases} 1 & |\mathcal{S}| \leq 1 \;\forall\; (\mathcal{S}, \mathcal{T}) \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

to allow assignment of multiple shots to the same sentence, but prevent one shot from being assigned to multiple sentences.

We now discuss alternatives to solve Eq. 7 and assign shots to sentences. We propose three different methods as a baseline and two efficient strategies based on dynamic programming.

4.1 Diagonal Prior

The simplest strategy for the alignment is to equally distribute the shots to sentences. Note that, we do *not* rely on any additional cues (character identities or subtitles) in this method. We assign $n = N_T/N_S$ shots to each sentence (assuming $N_S < N_T$) where N_T is the total number of shots in the video and N_S is the number of sentences in the plot. In general, any shot t_j is assigned to sentence s_i such that $i = \lceil j/n \rceil$.

The above assignment can be interpreted as setting $f(\cdot, \cdot) := 1$ and using a Gaussian distributed prior

$$P(s_i, t_j) \propto \exp\left[-\frac{(j - \mu_i)^2}{2\sigma^2} \right] \quad . \tag{9}$$

where $\mu_i = (i - \frac{1}{2}) \cdot n$. We empirically set $\sigma = n$ and keep $g(\cdot)$ as in Eq. 8 thus restricting each shot to be assigned to only one sentence. Fig. 2 (DIAG) shows an example of the resulting assignment obtained on one of the episodes in our data set. Note how the assignment is restricted to the diagonal.

4.2 Bow-shaped Prior

While the above is a simple prior with no information, it assumes that the story is equally important all through the episode. Typically, this is not the case as most stories tend to have a climax at the end. We observe that the presence of a climax causes the authors of the plot summaries to spend more sentences describing the end of the video (story) rather than the beginning or midsection. This is particularly true for movies where a lot of material can be interpreted as filler content.

To analyze the effect of this behavior, we incorporate a *bow-shaped prior* that shows smooth deviations from the diagonal. We model the shape of a bow using a *rational quadratic Bézier curve* [34] which can be

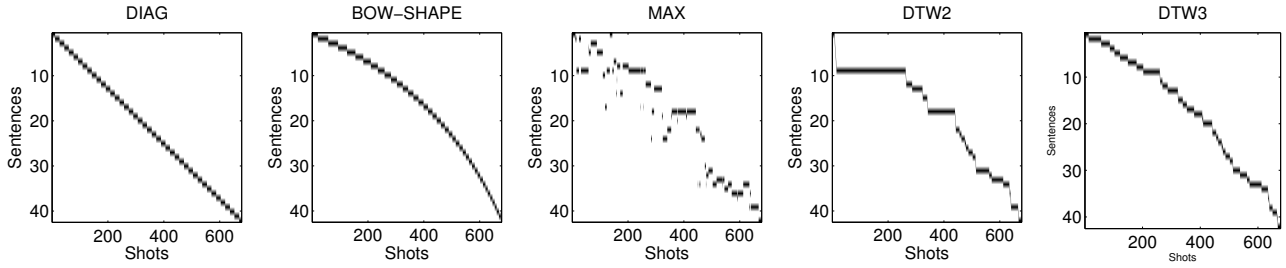

Fig. 2 From left to right: DIAG, BOW-SHAPE, MAX, DTW2, and DTW3 alignments for BF-01. The alignment accuracies are 2.8%, 17.7%, 11.6%, 30.9%, and 40.8% respectively.

used to parameterize conic sections given fixed start and end control points. The point which influences the amount of deviation is parameterized by $\gamma \in [0, 1]$ and is obtained as

$$P_{start} = [1, 1] \tag{10}$$
$$P_{end} = [N_T, N_S] \tag{11}$$
$$P_{\gamma} = [\gamma \cdot N_T,\ (1-\gamma) \cdot N_S] \tag{12}$$

While $\gamma = 0.5$ corresponds to the diagonal, $\gamma > 0.5$ creates a bow-shape and is used in our work to model the climactic nature of stories. Fig. 2 (BOW-SHAPE) shows an example alignment obtained by using a Bézier curve with $\gamma = 0.7$.

4.3 Max Similarity

As our final baseline, we use the information provided by the similarity cues – character identities and subtitles. From this method onwards, we also include the prior (Eq. 9) with the similarity scores $f(s_i, t_j)$. The goal of this method is to maximize the joint similarity of Eq. 7 while assigning one shot to only one sentence, thus fulfilling the global constraints (Eq. 8). All shots are treated independent to each other and shot t_j is assigned to a sentence s_i such that

$$i = \arg\max_i f(s_i, t_j) \cdot P(s_i, t_j) \tag{13}$$

Fig. 2 (MAX) shows the result of Max-Similarity. Note that as shots are treated independent from one another, we get an unlikely scenario of assigning neighboring shots to distant sentences. Nevertheless, the prior (Eq. 9) restricts the assignment to the major diagonal.

4.4 Max Similarity with Temporal Consistency

While Max Similarity actually achieves the maximum similarity score possible, we see that the resulting alignment is not natural. The assumption that shots are independent from one another causes problems.

In order to make the alignment temporally consistent, we restrict the assignment of a shot t_{j+1} to the same sentence s_i as t_j or to the next sentence s_{i+1}.

$$g(\mathcal{M}) = \begin{cases} 1 & |\mathcal{S}| \leq 1\ \forall\ (\mathcal{S}, \mathcal{T}) \in \mathcal{M} \text{ and} \\ & i \leq m \leq (i+1)\ \forall\ (s_i, t_j), (s_m, t_{j+1}) \in \mathcal{M} \\ 0 & \text{otherwise .} \end{cases} \tag{14}$$

We propose to use dynamic programming, specifically a modified version of the Dynamic Time Warping (DTW) algorithm [28] to perform the optimization efficiently. For simplicity, let us consider the similarity function $f(\cdot, \cdot)$ as a matrix of size $N_S \times N_T$, for N_S sentences and N_T shots. Each element of the matrix represents the similarity between one shot to one sentence. We enforce temporal consistency constraint Eq. 14 by allowing only two paths to arrive at any point on the DTW grid. A shot under consideration is assigned to either (i) the same sentence as the previous shot; or (ii) to the subsequent sentence. An example of such a grid is presented in Fig. 3 (left).

We construct a matrix D to store the scores obtained via exploration of all possibilities in the forward computation pass. The recursive update rules for the elements of D are

$$D(i, j) = \max \begin{cases} D(i, j-1) + f(s_i, t_j) \\ D(i-1, j-1) + f(s_i, t_j) \end{cases} \tag{15}$$

The highest scoring path is obtained via backtracking and corresponds to the optimal assignment $\mathcal{M}^*$ of shots to sentences. The backtracking starts at the last node in the grid $D(N_S+1, N_T+1)$. The computational complexity of this algorithm is in $\mathcal{O}(N_S N_T)$.

As this method uses a two-dimensional matrix as its grid, we label it as DTW2. Fig. 2 (DTW2) shows an example of the resulting alignment.

4.5 Regularized Max Similarity with Temporal Consistency

While DTW2 solves the problem of temporally consistent assignments, it says nothing about the number of shots that can be assigned to a sentence. Specially in cases when a sentence contains a large number of names, we observe that the similarity scores $f(s_i, \cdot)$ are consistently high. This can lead to a large number of shots being assigned to the same sentence (see Fig. 2 (DTW2)).

To prevent this erratic behavior, we propose an extension to the DTW2 algorithm. We introduce a decay factor α_k which decays as the number of shots assigned to the sentence increases. Empirically, we set a limit on the maximum number of shots that can be assigned to a sentence as $z = 5n$, where $n = N_T/N_S$ the average number of shots assigned to any sentence. The weights of the decay factor are computed as

$$\alpha_k = 1 - \left(\frac{k-1}{z}\right)^2, \quad k = 1, \ldots, z \tag{16}$$

The above addition can still be formulated as a dynamic programming problem and thus allows efficient solution. We incorporate the decay factor by extending our scores matrix D (from Sec. 4.4) by a third dimension $k = 1, \ldots, z$ to hold all possible paths. Now, when a shot is added to the same sentence, we not only traverse right, but also increase the depth by one level thus automatically counting the number of shots assigned to the current sentence. The update equation is

$$D(i, j, k) = D(i, j-1, k-1) + \alpha_k f(s_i, t_j), \quad \forall k > 1 \tag{17}$$

and assigning a shot to a new sentence resets the depth, setting $k = 1$

$$D(i, j, 1) = \max_{k=1,\ldots,z} D(i-1, j-1, k) + f(s_i, t_j) \; . \tag{18}$$

Note that we can arrive to a new sentence from any depth k as we do not know beforehand the ideal number of shots for the previous sentence.

Similar to DTW2, we compute the forward matrix D and then backtrack starting at the best depth level on the last node of the grid $\max_k D(N_S, N_T, k)$. We call this method DTW3 as it uses a three-dimensional matrix. The computational complexity of DTW3 is in $\mathcal{O}(N_S N_T z)$[3]. Fig. 2 (DTW3) shows an example of the resulting assignment. In contrast to DTW2, we see that DTW3 does not assign a large number of shots to one sentence.

[3] For $z \sim 100$, $N_S \sim 40$ and $N_T \sim 700$ DTW3 takes a couple of minutes to solve with our unoptimized Matlab implementation.

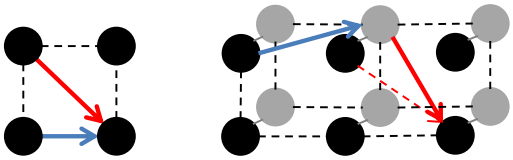

Fig. 3 LEFT: DTW2 (Sec. 4.4) valid paths. Blue/light indicates assignment of shot to the same sentence; red/dark indicates assignment to a new sentence. RIGHT: DTW3 (Sec. 4.5) valid paths. We represent the second depth layer in gray. Assigning a shot to the same sentence now also changes depth, and new sentences always start at the topmost layer.

4.6 Alignment Evaluation Measure

We propose a simple measure to evaluate the task of shot-sentence alignment. For a given shot t_j, we represent the corresponding aligned sentence through our automatic alignment method as $\mathcal{A}(t_j)$. Let $\mathcal{G}(t_j)$ be the ground truth sentence as provided in the alignment by the annotators. We measure the alignment accuracy as the fraction of shots that are correctly assigned by the automatic alignment to the ground truth annotation sentence:

$$ACC = \frac{1}{N_T} \sum_j \mathbb{1}\{\mathcal{A}(t_j) = \mathcal{G}(t_j)\} \; . \tag{19}$$

5 Story-based Search

We use the alignment between shots and sentences as the intermediate step to bridge the gap between textual and audio-visual representation of the story. Specifically, the story-based video retrieval problem is reduced to a text query in plot synopses.

We use Whoosh 2.5.4 [4], a full text indexing and search library to search within plot synopses. We index individual and groups of two and three sentences taken at a time as independent documents. The grouping of sentences into documents helps search for events which span a long duration; or search among sentences that only have pronoun references (*e.g. He slips away in his mist form.*) and do not hold sufficient information on their own. We use the BM25F [49] algorithm to generate a ranked list for document retrieval.

5.1 Evaluation measures

Motivated from a user perspective, we use two measures to evaluate the performance of our retrieval scheme.
(i) *top 5* (T_5): as a binary 0-1 answer, indicates whether the sentences belonging to the top 5 retrieved documents contain the story for which the user queries. This

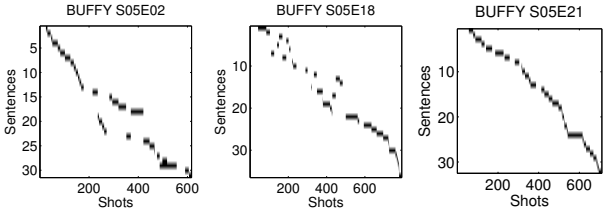

Fig. 4 Ground truth shot-sentence alignment annotations for multiple episodes: BF-02, BF-18 and BF-21.

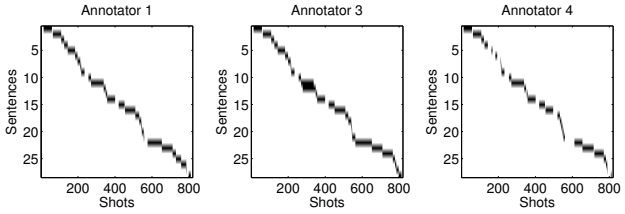

Fig. 5 Variation in ground truth annotations obtained from different annotators for BF-03. The Fleiss κ inter-rater agreement is 0.701 indicating substantial agreement.

measure essentially evaluates the quality of text search within plot synopsis.
(ii) *time difference* (t_Δ) / *overlap* (o_{IoU}): For the set of retrieved shots corresponding to the sentence, we have two possibilities. t_Δ counts the difference in time between the ground truth position of the story event and the set of returned shots (smaller is better). A desirable alignment results in an overlap between the ground truth shots and the retrieved shots in which case $t_\Delta = 0$ and the amount of overlap in time o_{IoU} is computed by an intersection-over-union (IoU) of the two time segments.

6 Evaluation

In this section, we evaluate the proposed algorithms to align shots to sentences and show that DTW3 provides best alignment, while a simple fusion of the subtitle and character identity cues works quite well. We also perform experiments on story retrieval and demonstrate encouraging performance.

6.1 Experimental Setup

6.1.1 Data set

To the best of our knowledge, searching through the storyline of a TV series is a new task. Thus, we build a new data set to evaluate our approach. The data consists of the complete season 5 of the TV series Buffy the Vampire Slayer (BF), a total of 22 episodes each ranging from 40-45 minutes. The series can be categorized as supernatural fantasy and contains a mixture of action, horror, comedy, and drama. The episodes follow a serialized format, *i.e.* each episode is a self-contained story contributing to a larger storyline which culminates at the season finale. Each episode contains about 720 shots on average (ranging from 538-940) and has a corresponding plot synopsis on Wikipedia which contains 36 sentences on average, (varying from 22-54). The data set is publicly available [1].

6.1.2 Ground truth Annotations

The problem of shot to sentence alignment is quite subjective and different people may have varied opinions on the shots which correspond to a sentence. We analyze the subjective nature of our problem by obtaining alignment labels from multiple people for a subset of our data (first 4 episodes BF-01 to BF-04).

For the human evaluation experiment, we gather labels from 4 annotators all in the range of 20-30 years. The annotators were asked to look at the entire video and select a begin and end shot for every sentence based entirely on the story conveyed in the corresponding text and video. To minimize bias, they were kept uninformed of the methods used for the alignment process.

Experiments on the complete data set are evaluated against one primary annotator who provided shot to sentence assignments for all 22 episodes. Fig. 4 shows ground truth annotations by this annotator for BF-02, BF-18 and BF-21. The graphs show the alignment path, *i.e.* the assignment of shots (on x-axis) to sentences (on y-axis).

From the resulting annotations, we derive some interesting properties:

1. Not all shots need to be assigned to a sentence. This reflects the idea that the plot synopses already form a type of summary;
2. The video and plot need not follow sequentially. The plot authors typically describe different storylines in separate paragraphs, while the episode editors tend to interweave them (*e.g.* BF-02, BF-18);
3. Although rare, multiple sentences are used to describe the same set of shots (BF-02, sentence 28-29).

While some of the above points violate our assumptions (Eq. 14), they occur very rarely and are ignored in the scope of the current work. However, note that our general formulation Eq. 7 allows for all of the above.

6.1.3 Cross-annotator Variation

We compare the alignment labels provided by different annotators in Fig. 5. While the overall structure

Table 1 Comparison of alignment accuracy against various alignment techniques.

Method		BF-01	BF-02	BF-03	BF-04
Fleiss κ		0.80	0.83	0.70	0.70
Human Accuracy		81.5	86.4	77.5	72.8
Diagonal Prior		2.9	23.8	27.9	8.8
Bow-shaped Prior		40.6	15.3	24.0	25.2
Character ID	MAX	11.6	30.9	23.6	19.1
Character ID	DTW2	9.4	35.0	18.7	28.4
Character ID	DTW3	**42.2**	43.8	**40.4**	**40.3**
Subtitles	DTW3	20.4	**48.4**	35.3	30.1
Char-ID+Subt.	DTW3	40.8	**51.3**	**41.4**	**47.6**

looks quite similar, the major differences include assignment of shots to neighboring sentences, skipping of shots, *etc*. We analyze the inter-rater agreement using the Fleiss κ [15] score by considering each shot as a *data sample* and the assigned sentence as a *category label*. To assist scoring, we introduce a *null* category which collects all shots that are not assigned to any sentence.

The κ scores for the four episodes are presented in Table 1 (row 1) and indicate substantial agreement (0.61-0.80) to almost perfect agreement (0.81-1.00).

We also compare annotators using our alignment evaluation measure as a "Human Accuracy" score by averaging the alignment accuracies obtained by taking pairs of annotators at a time. Presented in Table 1 (row 2), the human accuracy acts as an upper bound for any automatic method.

6.1.4 Queries for Story-based Retrieval

To evaluate the performance of story-based retrieval, we collect a total of 62 queries related to story events through the complete season. We obtain a fairly equal spread with 2 to 5 queries per episode. To reduce bias, a portion of the queries are obtained from a fan forum (`http://www.buffy-boards.com`) based on the TV series *Buffy the Vampire Slayer*, while the others are contributed by the annotators of the alignment. The annotators were instructed to only look at the video and not the plot synopsis, while creating additional queries.

Along with the queries, we include ground truth information such as the episode number in which the story appears and the time duration (correct up to 5 seconds) during which the plot unravels. The queries and the annotations are made available [1] for future comparison.

6.2 Quality of Character Identity Cues

Prior to analysis of the alignment performance, we briefly assess the performance of our character identity cue extraction methods (Sec. 3.2) both in the text and video domain.

Plot Synopses A plot synopsis in our data set contains on average 80 named references. We compare our automatic name extraction (including coreference resolution) against human annotations – a list of names inferred by reading the plot only and obtain a recall of 73% at a precision of 82%.

The main errors we encounter with the method can be attributed to: (i) inability to resolve plural pronoun references (*e.g. they*); and (ii) named references for characters who are referred to, but are not visible on-screen. For example, in *Riley asks Spike about Dracula*, it is clear to a human that *Dracula* does not appear while the automatic detection creates errors.

Videos An episode in our data set is about 40 minutes long and contains an average of 950 face tracks. Our identification labels each track with a name (among 59 characters) with an average accuracy of 63%. While this seems quite low, we show that the alignment is able to cope with errors. We also show that using ground truth person identification does not have a large influence.

6.3 Alignment Performance

Fig. 2 illustrates the alignments obtained from different methods. The diagonal prior (DIAG) equally distributes shots to sentences while the bow-shaped prior accounts for climax in an episode (BOW-SHAPE). The max similarity based method (MAX) achieves maximum joint similarity (Eq. 7) however assumes that shots are independent. The allowed transitions in dynamic programming based methods link shots together (DTW2) and finally DTW3 constrains the number of shots assigned to a sentence.

6.3.1 Alignment Methods

We evaluate methods discussed in Sec. 4 in combination with the two different cues – subtitles and character identities – and present the results in Table 1. The alignment accuracy is averaged across the annotators.

We report bow-shaped prior scores for the best choice of parameter γ over a grid search. Note how the prior can sometimes perform quite well and not all episodes favor the bow-shaped prior over the diagonal.

Table 2 Alignment accuracy on all episodes. The highest accuracy is highlighted in bold, while the second-best is italicized. The fusion (last column) performs best in most episodes, while character identities typically perform better than subtitles.

Episode	Diagonal Prior	Subtitles DTW3	Character ID DTW3	Char-ID+Subt. DTW3
BF-01	2.80	21.39	*40.85*	**42.77**
BF-02	20.29	*41.88*	39.12	**48.05**
BF-03	27.93	31.71	*32.32*	**32.68**
BF-04	4.20	24.37	*37.81*	**42.16**
BF-05	4.30	39.85	*45.33*	**51.11**
BF-06	7.65	33.02	*34.36*	**35.17**
BF-07	12.37	*52.15*	31.06	**55.43**
BF-08	12.73	*39.67*	36.69	**42.98**
BF-09	4.67	40.21	*40.96*	**48.80**
BF-10	5.71	*45.35*	43.23	**50.73**
BF-11	4.26	**50.73**	45.14	*49.80*
BF-12	9.54	41.91	*45.67*	**55.96**
BF-13	5.69	37.29	*48.67*	**61.62**
BF-14	1.89	*46.14*	21.27	**51.97**
BF-15	20.29	45.89	*57.56*	**60.34**
BF-16	9.66	27.70	*43.31*	**49.63**
BF-17	12.76	57.34	*64.69*	**69.93**
BF-18	6.06	27.27	*38.13*	**39.77**
BF-19	16.35	32.97	*54.59*	**62.16**
BF-20	10.00	19.79	*38.94*	**39.79**
BF-21	2.54	13.94	*34.51*	**51.83**
BF-22	20.75	**43.38**	31.54	*38.92*
Average	10.11	37.00	*41.17*	**49.16**

The DTW3 alignment with character identities outperforms the other methods and priors. As compared against character identities, subtitles are a weaker cue since the matrix of similarities $f_{subtt}(\cdot,\cdot)$ is very sparse. This results in lower accuracy. Nevertheless, fusion of the two cues (see Sec. 3.4) provides complementary information for the alignment. To account for the sparsity, we emphasize the subtitles and empirically set $\alpha = 2$.

6.3.2 Complete Season Evaluation

We present alignment results on the entire season in Table 2. The diagonal prior performs poorly at an alignment accuracy of 10.11%. On the other hand, the bow-shaped prior (we empirically determine $\gamma = 0.64$ for all episodes) shows better performance on average at 14.27%. *Character ID* with DTW3 outperforms *Subtitles* as a cue in most episodes (15 of 22) and their fusion produces the best result (in 20 of 22 episodes) at an average of 49.6%.

6.3.3 Relaxed Evaluation Metric

Our evaluation criteria is very strict and assignment of a shot to neighboring sentences is scored as an error. However, in practice, we often see that neighboring sentences typically discuss the same storyline. This is not true only when the sentences stem from two different paragraphs (and thus different storylines).

If we use a relaxed metric which allows alignment within ± 1 sentence, the alignment accuracy with cue fusion and DTW3 goes up to 71% (from 49%). Note that shots which are not assigned to any sentence in the ground truth are still counted as errors.

The impact with respect to retrieval is specially interesting. As the documents are composed of not only one, but two or three sentences taken at a time, it is likely that the queried event is found in a composite document. In such a case, an alignment within ± 1 sentence is acceptable. Nevertheless, note that the alignment evaluation metric (relaxed or otherwise) does not influence the retrieval performance.

6.3.4 Impact of Automatic Character Identification

To assess the quality of our face detection and tracking, we compute the Multiple Object Tracking Accuracy (MOTA) [7] score which takes into account false positives, missed detections and track switches. As labeling a face bounding box for every frame for all videos is a very time and cost intensive task, we evaluate performance on every 10th frame of the first 6 episodes of the season. The MOTA score averaged across all 6 episodes is 69.72%.

We also obtain 85.63% track recall (the number of tracks among ground truth which were detected); and a 88.73% track precision (the number of tracks which are actually faces and not false positive detections).

Further, our automatic face-based character identification scheme can tag a face track at an accuracy of 63%. When we use ground truth identity information, our alignment using DTW3 and identities goes up to 47.2%, about 6% higher. However, after fusion with subtitles, the alignment based on ground truth identities shows only a minor improvement of 2.7% to achieve 51.9% alignment accuracy. We can conclude that character identification is not the limiting factor, and given the current state-of-the-art systems, we achieve decent performance.

6.3.5 Qualitative results.

Fig. 6 presents a sample visualization of the alignment performance from the first episode of our data set. We visualize three sentences 27 to 29 and the set of corresponding shots.
(i) *BF-1:27* Note how the interaction between Buffy and Dracula is captured in the shots and character identification helps to assign shots 484–491 correctly to sentence 27.

Table 3 Performance of story-based retrieval on selected queries from the data set. E01:m35-36 means minutes 35-36 of episode 1. (33) indicates sentence number 33.

#	Query	Ground Truth		Retrieval		Time
		Location	Sentence	top 5	Sentence	deviation
1	Buffy fights Dracula	E01:m35-36	(33) Buffy and Dracula fight in a vicious battle.	✓	E01 (33)	$o_{IoU} = 10\%$
2	Toth's spell splits Xander into two personalities	E03:m11-12	(7) The demon hits Xander with light from a rod ...	✗	–	–
3	Monk tells Buffy that Dawn is the key	E05:m36-39	(34) He tells her that the key is a collection of energy put in human form, Dawn's form.	✓	E05 (34-35)	$o_{IoU} = 31\%$
4	A Queller demon attacks Joyce	E09:m32-33	(30) In Joyce's room, the demon falls from the ceiling ...	✓	E09 (28-30)	$o_{IoU} = 12\%$
5	Willow summons Olaf the troll	E11:m18-19	(17) Willow starts a spell, but Anya interrupts it ... (18) Accidentally, the spell calls forth a giant troll.	✗	–	–
6	Willow teleports Glory away	E13:m39-39	(34) ... before Willow and Tara perform a spell to teleport Glory somewhere else.	✓	E13 (34)	$o_{IoU} = 63\%$
7	Angel and Buffy in the graveyard	E17:m14-18	(13) At the graveyard, Angel does his best to comfort Buffy when she ...	✓	E17 (13-14)	$o_{IoU} = 61\%$
8	Glory sucks Tara's mind	E19:m24-27	(15) Protecting Dawn, Tara refuses, and Glory drains Tara's mind of sanity.	✓	E19 (14-15)	$o_{IoU} = 74\%$
9	Xander proposes Anya	E22:m16-19	(6) Xander proposes to Anya	✓	E22 (6)	$t_{\Delta} = 2m44s$

(ii) *BF-1:28* While the ground truth for this sentence ranges from shots 496–500, we assign shots 492–501 to this sentence. Shots 492–495 are assigned wrongly to sentence 28 and can be attributed to the fact that the character name *Dracula* is mentioned in the sentence and appears on screen. Shots 496–500 show the interaction between *Riley* and *Xander*.
(iii) *BF-1:29* We see the character *Giles* being overpowered by the *Three Sisters*. Again identities play a major role in the alignment performance.

In general, we observe a similar pattern in the alignment results across the episodes. A majority of the errors can be attributed to people appearing in multiple sentences inducing boundary localization errors.

6.4 Retrieval Performance

As mentioned previously, we evaluate story-based retrieval on 62 queries. A subset of the queries with their results (including failure cases) is shown in Table 3.

Given a query, we first search for matching words through the plot synopsis and present a ranked list of documents (sentence groups). We call a returned result successful when we obtain a document that corresponds to the queried story. Of the 62 queries, 24 (38%) obtain correct results in the first position, 43 (69%) are within the top 5 (T_5) and 48 (77%) are within the top 10 displayed results. For 9 of 62 (15%) queries we are unable to find a relevant document in the plot synopsis. The median position of a successful result is 2. In the current scope of the work we do not use synonyms to augment the query. Thus, rephrasing the query can often help improve performance.

Once a document is selected, our alignment is able to return a set of shots corresponding to the sentences. Of 53 queries for which we find a successful document, for 40 (75%) queries the set of returned shots overlap with the ground truth annotation. The remainder 13 queries are located on average about 3 minutes away from the depiction of the story in the video. Note that, considering that we search for a semantic event in all 22 episodes, or 15 hours of video, a 3 minute deviation is a very small fraction (0.33%).

6.5 Discussion: Plot Nonlinearity

One of the main challenges our alignment method currently faces is the non-causal nature of the plot synopsis and videos. While most plot descriptions and video shots are fairly straightforward, the editing structure of some episodes contain a mixture of shots belonging to different storylines. On the other hand the plot synopsis presents different storylines as separate paragraphs.

A different, but related problem is when the author of the plot synopsis does not deem a particular section of shots to be noteworthy and skips them in the description. A detailed analysis of the annotations shows that

BF-1:27. **Dracula** talks to **Buffy** of all the things **he** will do for **her** while **she** struggles to regain control of **herself**.

BF-1:28. **Xander** tries to stop **Riley** from going after **Dracula**, but **Riley** knocks **him** out with one punch.

BF-1:29. **Giles** finds **himself** victim to the **Three Sisters** who effectively keep **him** distracted.

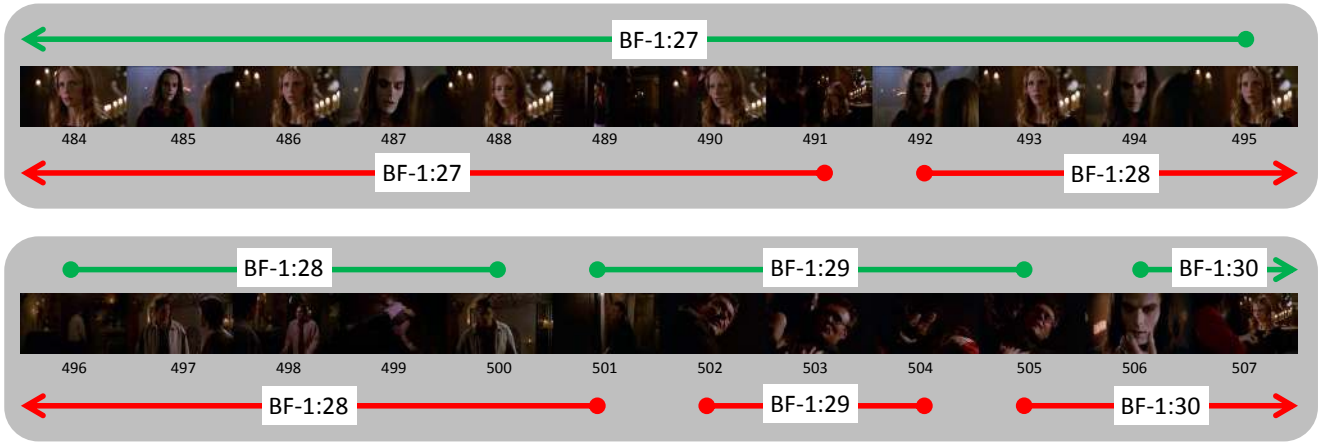

Fig. 6 Qualitative alignment results for 3 sentences from BF-01. The plot synopsis sentences (27 to 29) are presented at the top and the character references are highlighted for easy visualization. We show a list of 24 shots (484 to 507) with the shot number below the image and mark the ground truth alignment (green and above the shots) and predicted alignment (red and below the shots) as arrow segments. An arrow marker indicates continuation while ball markers signify termination. Refer to Sec. 6.3.5 for a discussion on the figure.

roughly 13% of shots are skipped, while our method forces every shot to be assigned to some sentence.

In case of movies, this problem is further accentuated as the video duration (and the number of shots) grows by about 3 times, while the number of sentences in plot synopses on Wikipedia do not scale accordingly. We will tackle these challenges in future work.

7 Conclusion

We present a novel problem of searching for story events within large collections of TV episodes. To facilitate the retrieval, we propose to align crowdsourced plot synopses with shots in the video. Such plot synopses or episode summaries are rich in content and are effective in capturing the story conveyed in the video. The alignment is formulated as an optimization problem and performed efficiently using dynamic programming. We evaluate the alignment against human annotations and show that 49% of the shots are assigned to the correct sentence. We also evaluate story-based retrieval on 15+ hours of video showing promising performance.

In the future, we intend to improve the alignment by using additional vision cues such as object detection, scene recognition and action recognition. An open research question is an efficient alignment for nonlinear video descriptions. We would also like to examine the alignment for other applications.

Acknowledgements This work was funded 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. Buffy Plot Synopsis Text-Video Alignment Data. https://cvhci.anthropomatik.kit.edu/~mtapaswi/projects/story_based_retrieval.html
2. NLP Toolbox. http://nlp.stanford.edu/software/
3. SubRip. http://en.wikipedia.org/wiki/SubRip
4. Whoosh - a Python full text indexing and search library. http://pypi.python.org/pypi/Whoosh
5. Alahari, K., Seguin, G., Sivic, J., Laptev, I.: Pose Estimation and Segmentation of People in 3D Movies. In: IEEE International Conference on Computer Vision (2013)
6. Bäuml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
7. Bernardin, K., Stiefelhagen, R.: Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. EURASIP Journal on Image and Video Processing pp. 1–10 (2008)
8. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
9. Bredin, H., Poignant, J., Tapaswi, M., Fortier, G., et al.: Fusion of speech, faces and text for person identification in TV broadcast. In: European Conference on Computer Vision Workshop on Information fusion in Computer Vision for Concept Recognition (2012)
10. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)

11. Cour, T., Sapp, B., Nagle, A., Taskar, B.: Talking Pictures : Temporal Grouping and Dialog-Supervised Person Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
12. Demarty, C.H., Penet, C., Scheld, M., Ionescu, B., Quang, V.L., Jiang, Y.G.: The MediaEval 2013 Affect Task: Violent Scenes Detection. In: Working Notes Proceedings of the MediaEval 2013 Workshop (2013)
13. Ercolessi, P., Bredin, H., Sénac, C.: StoViz: Story Visualization of TV Series. In: ACM Multimedia 2012 (2012)
14. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" - Automatic Naming of Characters in TV Video. In: British Machine Vision Conference (2006)
15. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971)
16. Freiburg, B., Kamps, J., Snoek, C.: Crowdsourcing Visual Detectors for Video Search. In: ACM Multimedia (2011)
17. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding Videos, Constructing Plots Learning a Visually Grounded Storyline Model from Annotated Videos Input. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
18. Habibian, A., Snoek, C.: Video2Sentence and Vice Versa. In: ACM Multimedia Demo (2013)
19. Jones, K.S.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation **28**, 11–21 (1972)
20. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-Scale Video Summarization Using Web-Image Priors. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
21. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning Realistic Human Actions from Movies. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
22. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Bruent, V., Boujemaa, N., Stentiford, F.I.: Video Copy Detection: a Comparative Study. In: ACM International Conference on Image and Video Retrieval (2007)
23. Law-To, J., Grefenstette, G., Gauvain, J.L.: VoxaleadNews: Robust Automatic Segmentation of Video into Browsable Content. In: ACM Multimedia (2009)
24. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In: Computational Natural Language Learning (2011)
25. Li, Y., Lee, S.H., Yeh, C.H., Kuo, C.C.: Techniques for Movie Content Analysis and Skimming. IEEE Signal Processing Magazine **23**(2), 79–89 (2006)
26. Liang, C., Xu, C., Cheng, J., Min, W., Lu, H.: Script-to-Movie : A Computational Framework for Story Movie Composition. IEEE Trans. on Multimedia **15**(2), 401–414 (2013)
27. Lin, D., Fidler, S., Kong, C., Urtasun, R.: Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
28. Myers, C.S., Rabiner, L.R.: A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition. Bell System Technical Journal (1981)
29. Nagel, H.: Steps toward a Cognitive Vision System. AI Magazine **25**(2), 31–50 (2004)
30. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured Learning of Human Interactions in TV Shows. IEEE Trans. on Pattern Analysis and Machine Intelligence (2012)
31. Peng, Y., Xiao, J.: Story-Based Retrieval by Learning and Measuring the Concept-Based and Content-Based Similarity. In: Advances in Multimedia Modeling (2010)
32. Poignant, J., Bredin, H., Le, V.B., Besacier, L., Barras, C., Quenot, G.: Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. In: Interspeech (2012)
33. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Trans. on Multimedia **7**(6), 1097–1105 (2005)
34. Rogers, D.F., Adams, J.A.: Mathematical Elements for Computer Graphics, 2 edn. McGraw-Hill (1990)
35. Sang, J., Xu, C.: Character-based Movie Summarization. In: ACM Multimedia (2010)
36. Sankar, P., Jawahar, C.V., Zisserman, A.: Subtitle-free Movie to Script Alignment. In: British Machine Vision Conference (2009)
37. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?" – Learning person specific classifiers from video. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
38. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: ACM Multimedia Information Retrieval (2006)
39. Snoek, C., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding Semantics to Detectors for Video Retrieval. IEEE Trans. on Multimedia **9**(5), 975–986 (2007)
40. Snoek, C., Worring, M.: Concept-Based Video Retrieval. Foundations and Trends in Information Retrieval **4**(2), 215–322 (2009)
41. Tan, C.C., Jiang, Y.G., Ngo, C.W.: Towards Textually Describing Complex Video Contents with Audio-Visual Concept Classifiers. In: ACM Multimedia (2011)
42. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: "Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
43. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Story-based Video Retrieval in TV series using Plot Synopses. In: ACM International Conference on Multimedia Retrieval (2014)
44. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: StoryGraphs: Visualizing Character Interactions as a Timeline. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
45. Tsoneva, T., Barbieri, M., Weda, H.: Automated summarization of narrative video on a semantic level. In: International Conference on Semantic Computing (2007)
46. Wang, X., Liu, Y., Wang, D., Wu, F.: Cross-media Topic Mining on Wikipedia. In: ACM Multimedia (2013)
47. Xu, C., Zhang, Y.F., Zhu, G., Rui, Y., Lu, H., Huang, Q.: Using Webcast Text for Semantic Event Detection in Broadcast Sports Video. IEEE Trans. on Multimedia **10**(7), 1342–1355 (2008)
48. Yusoff, Y., Christmas, W., Kittler, J.: A Study on Automatic Shot Change Detection. Multimedia Applications and Services (1998)
49. Zaragoza, H., Craswell, N., Taylor, M., Saria, S., Robertson, S.: Microsoft Cambridge at TREC-13: Web and HARD tracks. In: Proc. TREC (2004)