Book2Movie

Aligning Video scenes with Book chapters

Makarand Tapaswi, Martin Bäuml, Rainer Stiefelhagen

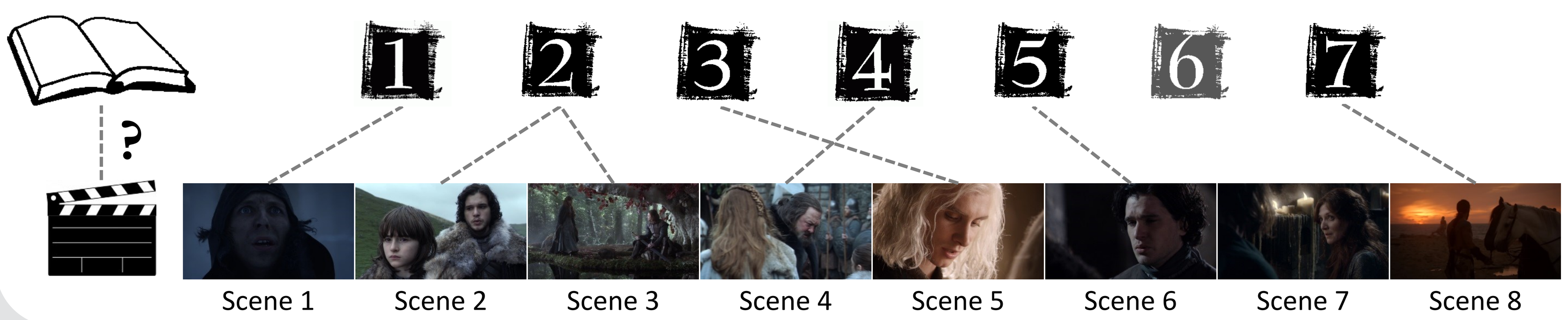
Computer Vision for Human Computer Interaction, Karlsruhe Institute of Technology, Germany

Highlights

- Joint analysis of source novels and their film and TV series adaptations
- New graph based model to align video scenes with book chapters, drops assumptions about sequential alignment
- Able to find differences between the adaptation and predict whether a scene was in the source book
- Extract rich textual paragraphs from the book which can be used to the describe the video

Motivation

TV series and films are often adapted from novels. Such adaptations are a large untapped resource to simultaneously improve story understanding for both vision and natural language processing. For example, descriptive text from the novels can be used to train video description models. Other applications include finding differences between the source and its adaptation.



Data set

Two diverse adaptations with shot level ground truth

Game of Thrones

- Many co-occurring stories
- Large cast list
- TV episodes, ~9h

Harry Potter

- Single, linear storyline
- Few central characters
- One movie, ~2h30m

Story characters and dialog parsing

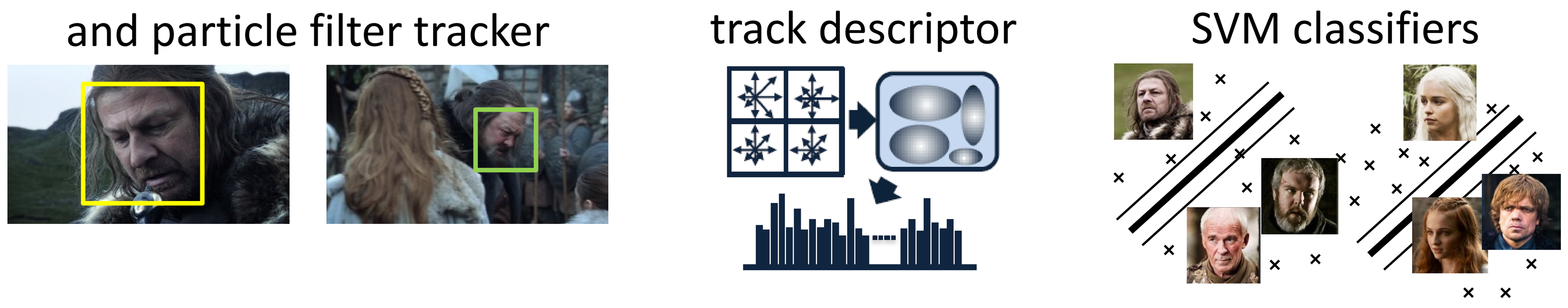
- Scene detection

- Characters in videos

Multi-pose face detector and particle filter tracker

Fisher Vector track descriptor

One-vs-all SVM classifiers

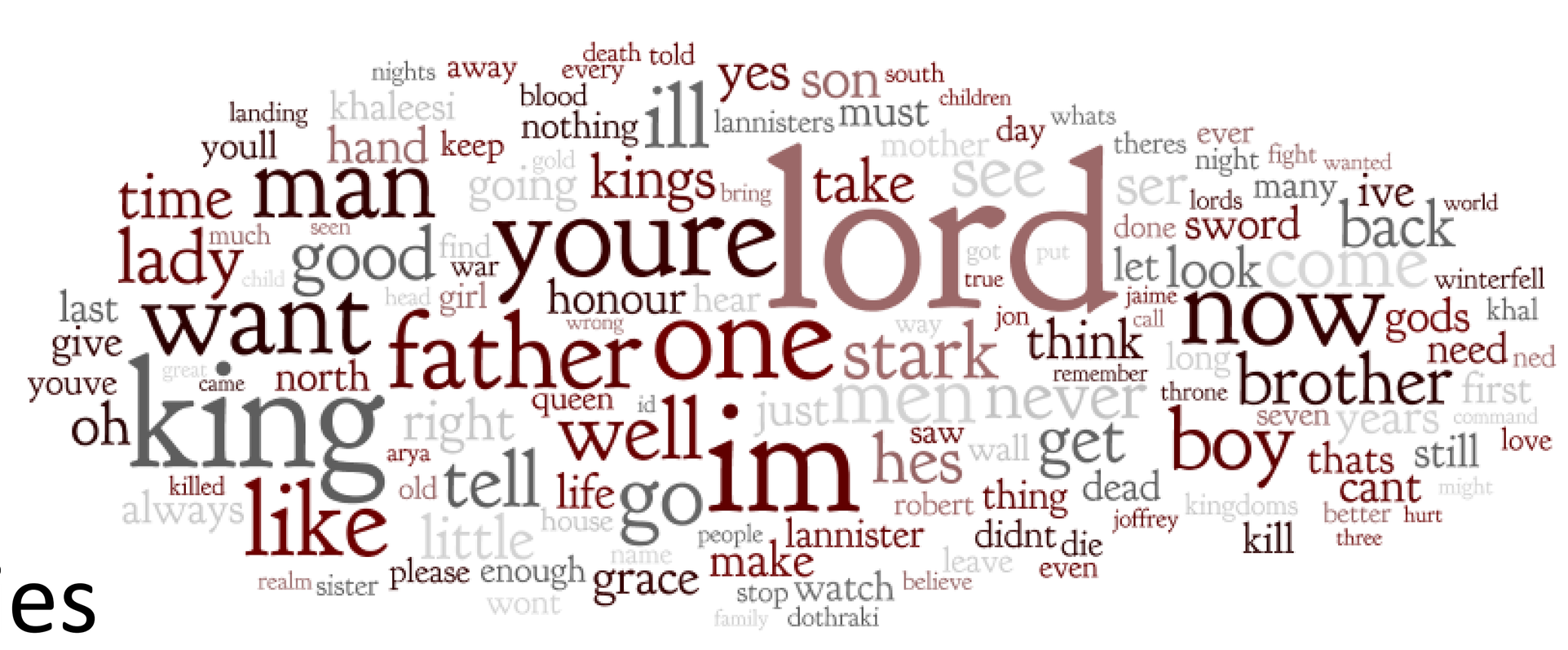


- Character mentions in book

full name > first name > alias, titles > last name

e.g. Eddard Stark > Eddard > Ned, Lord Stark > Stark

- Dialogs in videos CC directly from subtitles
- Dialogs in books quoted speech
- Word importance inverted term frequencies



Contact

{tapaswi,baeuml}@kit.edu

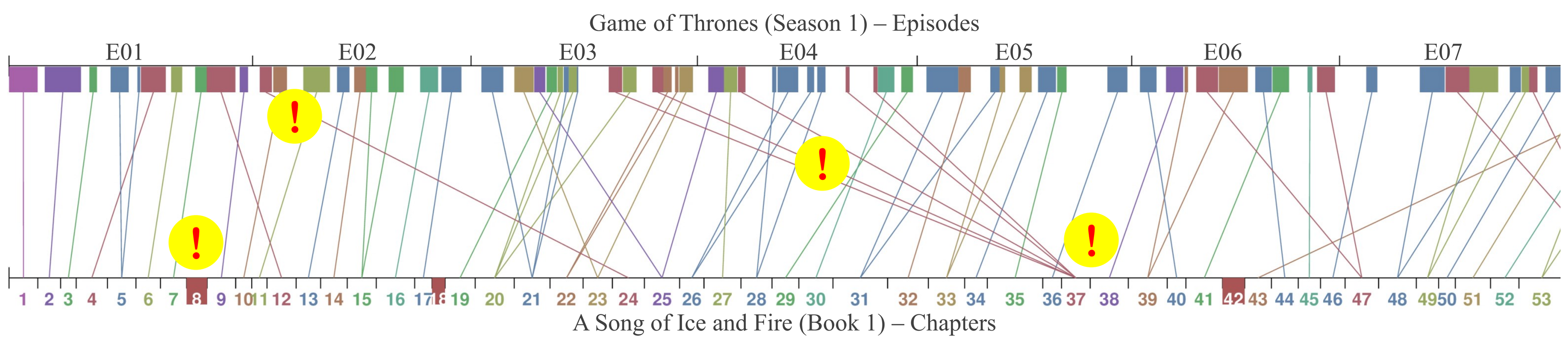
Project page (data)

http://cvhci.anthropomatik.kit.edu/projects/mma

Alignment

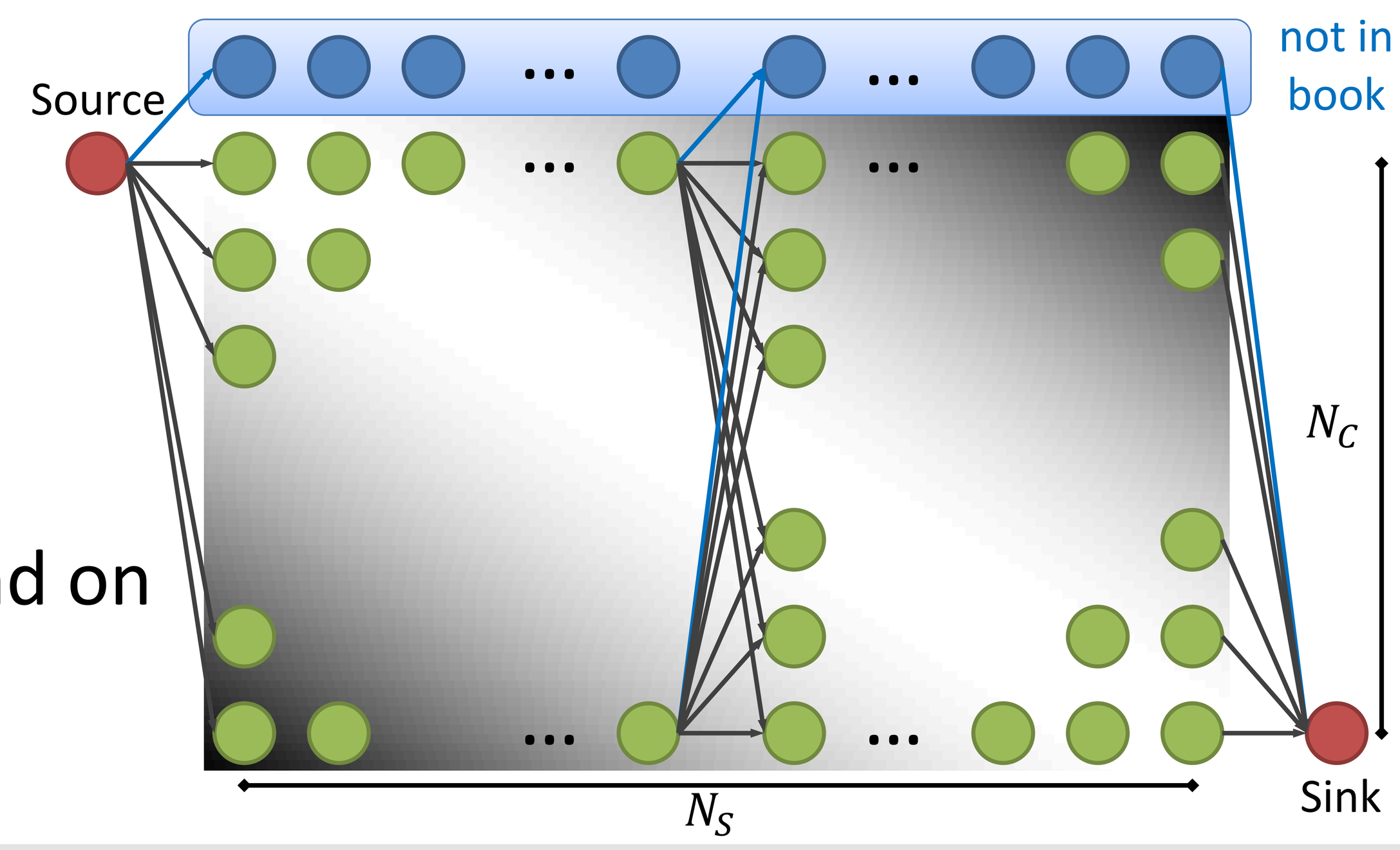
- Matching characters and dialog

- Difficulties in alignment



- Formulate as a shortest path problem
- Shortest path through the graph ⇔ best alignment

- Features
  - local prior
  - global prior
  - scenes not in book
  - jump chapters
- Edge weights depend on
  - character identities
  - matching dialogs



Evaluation

- Data set statistics

	VIDEO		BOOK		Face-ID	
	duration	#scenes	#chapter	#words	#charact.	id acc.
GOT	8h 58m	369	73	293k	95	67.6
HP	2h 32m	138	17	78k	46	72.3

- Alignment performance
  - upper bound
  - prior helps
  - dialogs strong
  - linear story lines DTW3 is good
  - adding ∅ helps

	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**	–	–

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

Mining rich descriptions

Ch7, P131
M 46m55s

Harry, who was starting to feel warm and sleepy, looked up at the High Table again. Professor Quirrell, in his absurd turban, was talking to a teacher with greasy black hair, a hooked nose, and sallow skin.

Ch23, P83
E03 50m37s

A slight man with a bald head and a great beak of a nose stepped out of the shadows, holding a pair of slender wooden swords. "Tomorrow you will be here at midday," He had an accent, the lilt of the Free Cities, Braavos perhaps, or Myr.

Acknowledgment: This work was funded by the German Research Foundation (DFG).