Automatic annotation of unique locations from video and text

Chris Engels[1]
Chris.Engels@esat.kuleuven.be
Koen Deschacht[2]
Koen.Deschacht@cs.kuleuven.be
Jan Hendrik Becker[1]
JanHendrik.Becker@esat.kuleuven.be
Tinne Tuytelaars[1]
Tinne.Tuytelaars@esat.kuleuven.be
Marie-Francine Moens[2]
Sien.Moens@cs.kuleuven.be
Luc Van Gool[13]
Luc.VanGool@esat.kuleuven.be

[1] ESAT-PSI
K.U. Leuven
[2] Department of Computer Science
K.U. Leuven
[3] Computer Vision Laboratory
BIWI/ETH Zürich

In this paper, we tackle the challenging problem of extracting information from unstructured text and exploiting this information to annotate an associated video. In particular, we develop a method that operates on a video with associated transcript, segments the video into scenes that are set in a specific location, and automatically annotates each scene with a textual label that provides a description of that location (*e.g.* "Joyce's living room").

We focus on action series, which present many challenges for automated location annotation. Most substantially, many locations are unique to specific episodes. Thus, any supervised approach that assumes annotations from other episodes is guaranteed to fail. Other episodes can only be used to learn global statistics and parameter settings. To generate the actual location labels, we only use the unstructured transcript of the episode being annotated. Other problems for visual recognition in this context include rapid camera switching and motions, nondescript or blurred backgrounds, and individuals appearing in multiple locations.

We provide an overview of our approach in Fig. 1. Given a video and associated transcript, we first approximately align the two using timing information from the subtitles. We segment the scenes using a cross-modal approach, which finds potential scene cuts in the text and shot cuts in the video, and then constructs the scenes using agglomerative clustering.

The transcripts describe many of the locations within each episode, but this information may be hidden within the text or not present for some scenes. The solution we propose is based on generating textual labels using a weighted topic mixture model. We first rely on an automatic detector to find likely location descriptions in the text. These detected descriptions are used to learn topic mixtures for each scene with Latent Dirichlet Allocation (LDA). Using this topic model allows us to use contextual information when labeling a location and to handle different wordings for the same location. Finally, we modify the topic distribution for each scene using a visual similarity measure based on Earth Mover's Distance; this step propagates labels to scenes lacking informative text.

Unlike supervised methods that output a unique class for a given input, our system generates a list of phrases and corresponding probabilities, with the most likely phrase being assigned as the label. Because these phrases are generated within the context of the transcript, multiple phrases could be considered valid. Therefore, creating a useful ground truth for automatic evaluation of the final result is nontrivial, so we instead rely on a qualitative evaluation to validate our approach.

We test our approach on four episodes of *Buffy the Vampire Slayer*, which provides a challenging validation for our system. We report qualitative results in Table 1 for three different settings: the full system as described above (*text+lda+vision*), without the vision (*text+lda*) and without LDA (*text*).

episode	text	text+lda	text+lda+vision
1	58.38% ± 4.88%	61.49% ± 4.88%	**67.62% ± 7.85%**
2	64.37% ± 8.23%	67.87% ± 7.91%	**75.72% ± 6.76%**
3	69.86% ± 3.60%	69.55% ± 5.96%	**71.43% ± 6.97%**
4	53.36% ± 4.11%	55.46% ± 4.81%	**63.02% ± 10.26%**
all	61.50% ± 8.24%	63.59% ± 8.11%	**69.45% ± 9.16%**

Table 1: Accuracy and standard deviation of the proposed systems.

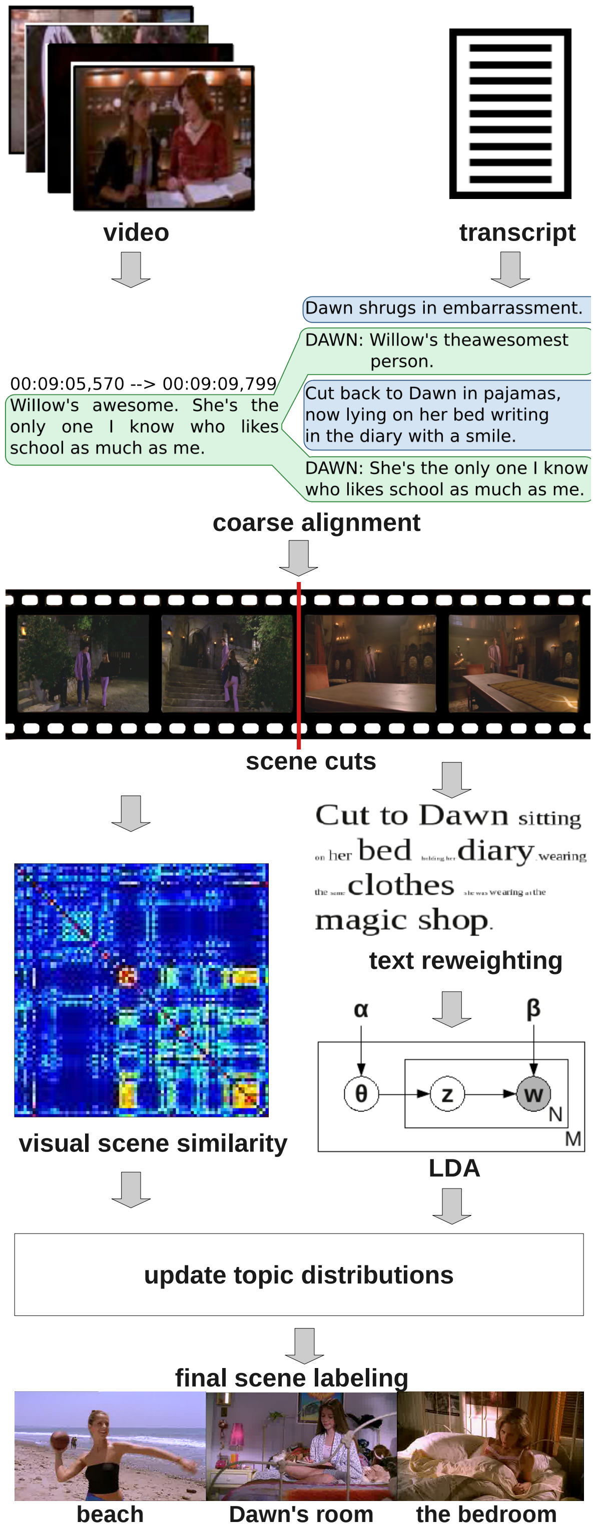

Figure 1: Overview of our approach.