Subtitle-free Movie to Script Alignment

Pramod Sankar K.[1]
pramod_sankar@research.iiit.ac.in
C. V. Jawahar[1]
jawahar@iiit.ac.in
Andrew Zisserman[2]
az@robots.ox.ac.uk

[1] Centre for Visual Information Technology
IIIT, Hyderabad, INDIA
http://cvit.iiit.ac.in

[2] Visual Geometry Group
University of Oxford, Oxford, UK
http://www.robots.ox.ac.uk/~vgg

Scripts of movies and TV videos, offer a number of possibilities for video understanding: they can provide supervisory information for identifying characters [2] or learning actions [3]; they enable text-based retrieval and search [4]; they enable a scene level organization of the video material [1], etc. The typical method of aligning a script (or transcript) with TV or movie videos, is by applying dynamic time warping with the subtitles, as introduced by Everingham *et al*. [2]. However, subtitles are not readily available for many old films, silent movies and non-European language videos. In this paper, our objective is the visual alignment between TV/movie videos and their scripts, *without* using subtitles. Achieving such an alignment increases the scope and applicability of script-based approaches to videos with little or no spoken dialogue. The challenge of the problem is in the comparison and matching of visual information of the video with descriptions given in the script. These descriptions generally involve objects and actions for which visual recognition is not yet mature.

Our approach is to combine several cues, both visual and auditory, which in themselves are not quite reliable, but when combined provide sufficiently strong constraints for a full alignment. We pose the problem as one of multi-state labelling of a sequence of shots, where the states for each shot correspond to the sentences of the script. For a correct alignment, each sentence should compete for the right shot to fall into. The voting of a shot should depend on common features that can be extracted from both the sentence and the shot. Towards this end, we extract three clues from each shot/sentence: $< Location, Face, Speech >$.

Location recognition helps in localizing sentences to the shots belonging to the given scene. We begin by locating *stock-shots* using a near-duplicate detection technique. This specifies the start of the scene. The location of each shot is then recognised using a Bag-of-Visual-Words [5] based SVM classifier. By combining the results from both, we obtain the temporal segmetation of the video.

It is more likely that a sentence belongs to a shot where the speaker of the sentence is visible. To ascertain the presence of the particular character in a shot, we detect and recognize faces across the shot. The detected faces are matched against a set of exemplars for each character. A Kernel-SVM using the min-min distance between face tracks is learnt for each person. The classifier with the best score gives the label for the faces in the shot.

Further evidence of sentence-shot correspondence could be obtained from speech recognition. We recognize the speech of each shot using a commercial speech recognizer. Though the actual recognition performance is quite poor, the limited matches do help in the alignment.

Independently, each of the audio-visual recognition modules are not accurate enough to align the movie with the script. We thus integrate all the clues into a local assignment cost $d(i,j)$ as

$$d(i,j) = \alpha_1 \cdot Cost_{Location}(i,j) + \alpha_2 \cdot Cost_{Face}(i,j) + \alpha_3 \cdot Cost_{Speech}(i,j),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The matrix of global alignment costs is denoted as $\mathscr{D}$. Our formulation lends itself to be solved using dynamic programming (DP). By backtracking the array $\mathscr{D}$, we recover the alignment between the sentences and shots.

Our approach was applied on episodes from the popular TV show *Seinfeld*, on clips from Charlie Chaplin's silent films Gold Rush and City Lights, and on Indian movies. For the Seinfeld videos, the weights α for each modality are learnt using two episodes as training data. We obtain an alignment accuracy of 74% using our technique. In comparison, a subtitle-based alignment achieves an accuracy of 91%.

Example results of annotation following the script alignment is shown in Figure 1 (above). In most error cases, the sentences are assigned within a few shots of the actual shot they belong to. Over our test data, the maximum distance of an erroneous assignment was five shots. In a video retrieval scenario, for a given textual query, we could provide a video segment consisting of multiple shots that would most likely contain the right answer. We thus achieve considerable retrieval performance, even though the precise alignment might be less accurate.

We further demonstrate the applicability of our approach by aligning scripts with silent movies of Charlie Chaplin (Figure 1 (below)), and on Indian movies. Inspite of the lack of subtitles in both cases, we were able to satisfactorily align the video with the dialogues and descriptions from the script.

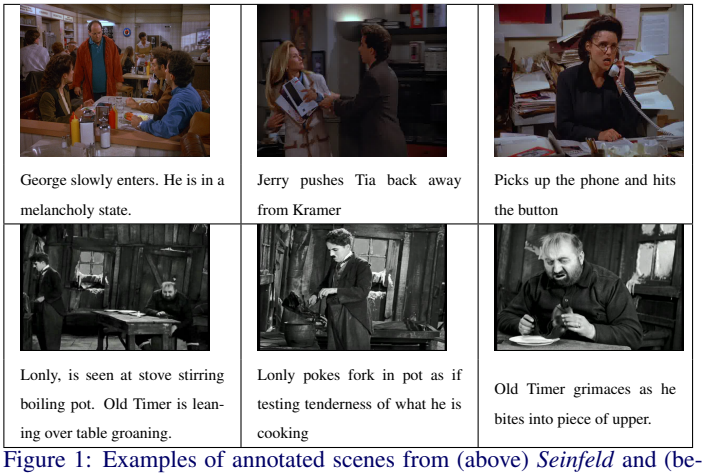

Figure 1: Examples of annotated scenes from (above) *Seinfeld* and (below) the *Gold Rush*.

We have presented a framework that can be extended as recognition of objects, actions and speech improves, so that more correspondences can be found between nouns/verbs in the script and the video.

[1] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Proc. ECCV*, 2008.

[2] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proc. BMVC*, 2006.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008.

[4] K. Pramod Sankar, S. Pandey, and C. V. Jawahar. Text driven temporal segmentation of cricket videos. In *Proc. ICVGIP*, 2006.

[5] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.