Segmentation and Annotation of Audiovisual Recordings Based on Automated Speech Recognition

Stephan Repp[1], Jörg Waitelonis[2], Harald Sack[2], and Christoph Meinel[1]

[1] Hasso-Plattner-Institut für Softwaresystemtechnik GmbH (HPI), P.O. Box 900460, D-14440 Potsdam, Germany
[2] Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-07743 Jena, Germany

Abstract. Searching multimedia data in particular audiovisual data is still a challenging task to fulfill. The number of digital video recordings has increased dramatically as recording technology has become more affordable and network infrastructure has become easy enough to provide download and streaming solutions. But, the accessibility and traceability of its content for further use is still rather limited. In our paper we are describing and evaluating a new approach to synchronizing auxiliary text-based material as, e. g. presentation slides with lecture video recordings. Our goal is to show that the tentative transliteration is sufficient for synchronization. Different approaches to synchronize textual material with deficient transliterations of lecture recordings are discussed and evaluated in this paper. Our evaluation data-set is based on different languages and various speakers' recordings.

1 Introduction

Audiovisual recordings in terms of streaming media or video podcasts (vodcasts) are increasingly used for live distance and on-demand lecturing by universities and distance learning institutions. Independent in/from time and place, learners have access to libraries of recorded lectures, often being organized as knowledge bases that offer their content in a well ordered and categorized manner. But, how can appropriate information be retrieved in a large lecture video data base in an efficient way? Manual segmentation of video lectures into smaller units, each segment related to a specific topic, is an accepted approach to find the desired piece of information [7,14,19,17,18].

Traditional multimedia retrieval based on feature extraction cannot be efficiently applied to lecture recordings. Lecture recordings are characterized by a homogeneous scene composition. Thus, image analysis of lecture videos fails even if the producer tries to loosen the scene with creative camera trajectory. A promising approach is based on using the audio layer of a lecture recording in order to get information about the lecture content. Unfortunately, most lecture recordings do not provide optimal sound quality and thus, also the effectiveness of automatic speech recognition (ASR) for the extraction of spoken words suffers even if a speaker-dependent system is used. The raw results of an untrained

H. Yin et al. (Eds.): IDEAL 2007, LNCS 4881, pp. 620–629, 2007.

ASR applied to lecture audio streams are not capable for an accurate indexing. Today, in lectures often text-based presentations such as MS Powerpoint or Portable Document Format (PDF) are used to support the teaching. The best alternative is thus to synchronize slides with presenter speech, which can be extracted using a speech to text engine.

Section 2 of the paper shows current approaches and outlines related work, while section 3 introduces the new segmentation method. Section 4 gives an evaluation of the algorithm and Section 5 concludes the paper with a short summary and an outlook on future work.

2 Related Work

Using speech recognition for indexing the lecture videos is an often used and evaluated procedure [10,7,14]. Text matching is a widely recognized method [12] too. The Text matching method has a direct impact on other fields such as genetic sequence alignment [11]. The main focus of these algorithms is to parse large strings on a block level and to find correct matches to short strings (gene). Much research has been focused on the area of intelligent meeting rooms and virtual classrooms [9,6]. They usually try to minimize error in the speech transcription or recognition stages. Chu and Chen presented an approach for cross-media synchronization [4]. They match audio recordings with text transliterations of the audio recordings based on dynamic programming methods [13]. It differs from our approach. In our case, the content of the presentation slides and the transliteration of audio recording do not match. Chu and Chen make use of explicitly encoded events for synchronization, instead of implicit automated synchronization.

Another non-analytic approach is to synchronize presentation slides by maintaining a log file during the presentation thus keeping track of slide changes. But, most available lecture recordings do neither support desktop recordings nor maintain a dedicated log file. In [19], they synchronize the book sections with word blocks of the transcript. Yamamoto use a sliding window system and a vector model. They measure the precision and the recall of the results. Chen and Heng [2] also propose a method for automatic synchronisation of speech transcripts with presentation slides. After matching the speech transcript with the slides, redundancy and noise are removed by a fitting procedure. Finally, transitions of slides are approximated by using a progressive polynomial fitting.

Our algorithm lies downstream of these related works, assuming speech to text transcription from an out-of-the box commercial speech recognition software. Further we show, how linear text segmentation algorithms [5,3] can be applied to segment lecture video recordings and to map presentation slides to a particular point of time in the video recording. Additionally we suggest a new algorithm, that deals with achieving global synchronization between the transcript and the presentation text. In the synchronization problem, our transcript input is expected to contain a large amount of error from the speech to text process. Our evaluation is divided into two stages: