AI & Soc
DOI 10.1007/s00146-013-0452-4

OPEN FORUM

Spoken query based word spotting in digitized Tamil documents

AN. Sigappi · S. Palanivel

Received: 22 October 2012 / Accepted: 25 February 2013

Abstract This paper presents an integrated approach to spot the spoken keywords in digitized Tamil documents by combining word image matching and spoken word recognition techniques. The work involves the segmentation of document images into words, creation of an index of keywords, and construction of word image hidden Markov model (HMM) and speech HMM for each keyword. The word image HMMs are constructed using seven dimensional profile and statistical moment features and used to recognize a segmented word image for possible inclusion of the keyword in the index. The spoken query word is recognized using the most likelihood of the speech HMMs using the 39 dimensional mel frequency cepstral coefficients derived from the speech samples of the keywords. The positional details of the search keyword obtained from the automatically updated index retrieve the relevant portion of text from the document during word spotting. The performance measures such as recall, precision, and F-measure are calculated for 40 test words from the four groups of literary documents to illustrate the ability of the proposed scheme and highlight its worthiness in the emerging multilingual information retrieval scenario.

Keywords Word spotting · Information retrieval · Segmentation · Mel frequency cepstral coefficients · Hidden Markov models

AN. Sigappi (✉) · S. Palanivel
Department of Computer Science and Engineering,
Annamalai University, Annamalainagar 608 002, India
e-mail: aucse_sigappi@yahoo.com

S. Palanivel
e-mail: spal_yughu@yahoo.com

1 Introduction

The widespread reach of Internet and digital storage media along with index based search and retrieval mechanisms facilitates easy access to information stored in digital formats. However, searches made on digitized documents, such as scanned images of handwritten and printed documents, continue to espouse challenges in view of the inherent difficulties that arise in providing the search word to the information retrieval system, absence of an index of keywords, and degradation in the quality of digitized documents (Likforman-Sulem et al. 2007; Leydier et al. 2009). Yet, another perspective to information retrieval that draws the attention of researchers in recent times is the deployment of voice based searches that offer a natural interface to the user and result in the elimination of keyword typing and/or keyword image selection for submission of the query input. Tamil is a classical, Dravidian language spoken largely by people living in the Indian state of Tamil Nadu and by a sizeable population living across the world. Tamil literature continues to evolve from the pre Christian era and contributes to the growth of Tamil language through poetry, prose, and grammar recorded for future reference in the form of stone inscriptions, palm leaf manuscripts, paper based manuscripts, machine printed documents, and of late as digital documents. Owing to the fact that traditional literature content is invariably referred, the steps taken to digitize ancient handwritten and machine printed documents is required to be supplemented and realize an effortless insight to digitized information.

A spoken word based search is a more natural and convenient form of gaining access to locate information present in digitized documents than using the word images from the index as search queries. With the steadily growing increase in the availability of information in regional

Published online: 21 March 2013