Textual-Content Based Classification of Bundles of Untranscribed of Manuscript Images

Jose Ramón Prieto, Vicente Bosch, Enrique Vidal
PRHLT Research Center
Universitat Politècnica de València, Spain
email: evidal@prhlt.upv.es

Carlos Alonso, M. Carmen Orcero, Lourdes Marquez
Centro de Arqueología Sunacuática
Insituto Andaluz del Patrimonio Histórico, Seviall, Spain
email: carlos.alonso.v@juntadeandalucia.es

***Abstract*—Content based classification of manuscripts is an important task that is generally performed in archives and libraries by experts with wealth of knowledge on the manuscripts contents. Unfortunately, many manuscript collections are so vast that it is not feasible to rely solely on experts to perform this task. Current approaches for textual-content based manuscript classification generally require the handwritten images to be first transcribed into text – but achieving sufficiently accurate transcripts is generally unfeasible for large sets of historical manuscripts. We propose a new approach to automatically perform this classification task which does not rely on any explicit image transcripts. It is based on "probabilistic indexing", a relatively novel technology which allows to effectively represent the intrinsic word level uncertainty generally exhibited by handwritten text images. We assess the performance of this approach on a large collection of complex manuscripts from the *Spanish Archivo General de Indias*, with promising results.**

I. Introduction

We consider the task of automatic classification of bundles or folders (hereafter called "documents") of manuscripts, according to their textual contents. We assume that the manuscripts of interest have been scanned into into high resolution digital images and the task consists in classifying a given document, that may range from a few to a few thousands of handwritten text images, into a predefined set of classes. Classes are associated with the topic or (semantic) content conveyed by the text written in the document images. When we say "classification of handwritten text images by their textual contents" it is advisable to avoid some frequent confusions.

First of all, this task is very different from what in the computer vision and image analysis literature is usually called "image classification" [1], [2], [3], where images are classified according to more or less global features related to colours, textures, shapes, etc. It is also very diferent from the task often referred to as "content-based image classification" [4], [5]. In a conventional content-based image classification task, images typically contain a few relatively large objects, such as mountains, animals, vehicles, persons, etc., out of a few tens (or maybe a few thousands) types of objects. In contrast, a typical text image contains several hundreds of small and detailed "objects" (i.e., words), out of several tens (or hundreds) of thousands "types" of different "objects" (i.e., different words of a natural language lexicon). For similar reasons, works such as [6], [7], where visual and text features are combined, are not comparable with the work here presented.

Another mix up which is worth avoiding is to relate the task considered here with what in the document analysis literature is often called "image document classification", where images of printed or handwritten text are classified according to more or less global features such as layout visual shape, type of script, writer (hand), etc. [8], [9], [10].

Instead, what we intend to do is similar to the time-honoured and well known task of *content based document classification*, which assumes the data are plaintext documents, rather than handwritten text image documents. Traditional examples, for which popular datasets are available, are *Twenty News Groups*, *Reuters*, *WebKB*, etc. [11], [12], [13].

For the task here considered (textual-content based handwritten text image document classification), the current commonly accepted wisdom is to split the process into two sequential stages. First a handwritten text recognition (HTR) system should be used to be transcribe the images into text and, second, traditional content based document classification methods can be applied to the resulting text documents.

This approach might work to some extent for simple manuscripts with uniform writing style and good quality images, where HTR can provide highly precise transcripts with over 90% word recognition accuracy [14]. It can of course work also for small-scale collections, where manual correction of HTR errors can be affordable.

But this is not an option for countless large historical collections of, say, hundreds of thousands images. Moreover, for many of these collections the best available HTR systems can only provide word recognition accuracies as low as 50-70% (e.g the ICDAR-2015 benchmark [14]). This is the case of the Carabela collection considered in this paper. It encompasses more than the 125 000 complex page images [15] and the average word recognition accuracy achieved in optimistic laboratory conditions is 65% [16], dropping to 46% when conditions are closer to real-world usage [15].

Clearly, for this kind of manuscript collections, the aforementioned two-stage idea is to be ruled out and new, holistic approaches should be devised. To the best of our knowledge, this is the first paper proposing, developing and assessing this kind of approaches on a large mansucript dataset – notwithstanding previous publications dealing with related problems and ideas, mainly aimed at printed text [17], [18], [19].