WebGT: An Interactive Web-based System for Historical Document Ground Truth Generation

Ofer Biller, Abedelkadir Asi, Klara Kedem and Jihad El-Sana
Computer Science Department
Ben-Gurion University of the Negev
Beer-Sheva, Israel
billero,abedas.el-sana,klara@cs.bgu.ac.il

Itshak Dinstein
Electrical Engineering Department
Ben-Gurion University of the Negev
Beer-Sheva, Israel
dinstein@ee.bgu.ac.il

Abstract—We present WebGT, the first web-based system to help users produce ground truth data for document images. This user-friendly software system helps historians and computer scientists collectively annotate historical documents. It supports real time collaboration among remote sites independent of the local operating system and also provides several novel semi-automatic tools that have proven effective for annotating degraded documents.

I. INTRODUCTION

There is a new age for computer analysis of digitized ancient documents. Researchers are developing algorithms to determine page layout, segment text blocks, illustrations, lines, words, letters, etc. Machine learning techniques have become an integral part of algorithms for document analysis. They rely heavily on ground truth data to generate reliable trained models. Moreover, ground truth availability enables researchers to employ quantitative comparisons among different methods. As a consequence, strengths and weaknesses of these methods can be revealed.

Several ground-truth generation systems have been suggested in the literature. These systems have the ability to work on different levels of granularity such as text lines, connected components, and individual pixels. Pink Panther [9], is one of the early tools for ground truth generation and document layout evaluation. This tool allows users to view a document image, zone it with simple mouse clicks, and specify information about each zone, e.g. the zone type, as well as the ordering relationships between zones. Zones can be defined using rectangles, constrained polygons, and arbitrary polygons. TRUEVIZA [5], developed by Lee and Kanungo, is a tool for visualizing and creating ground truth. It enables the users to generate ground truth for text lines, words and glyphs together with their textual content. The PerfectDoc [8] system has similar functionalities to the previous system, however, it incorporates advanced features such as multi page reading order and semantic annotations. PixLabeler [7] is a pixel-based system. It enables users to select groups of foreground pixels that are likely to be meaningful elements and label them with one of a pre-defined set of labels. But apart from different labels, no further data can be entered for elements. A Java based system, GEDI, was suggested by Doerman et al. [3]. This system provides a comprehensive set of ground truth generation tools for different layout analysis tasks. Among others, it enables users to define different types of zones, e.g., text and graphics zones, each of which may have a custom set of attributes. It offers a special functionality for text annotation which avoids drawing bounding boxes. Researchers can transcribe the zone content first and provide optional offsets among components later on. Fischer et al. [4] combined several algorithms and suggested an off-line semi-automatic (see details below) framework which is used for ground truth creation. Their method addresses ground truth creation for historical documents. Recently, Aletheia, a scalable and robust system was suggested by Clausner et al. [2]. It is completely flexible as it supports top-down as well as bottom-up ground-truth generation schemes. In top-down scheme users select complex structures, such as text regions, in the documents and then refine this selection with split and shrink tools. In the bottom-up scheme, users aggregate basic elements to high-level entities. This system enables users to define region outlines using polygons and rectangles. The systems described above cover different aspects of ground truth generation, and most of them have no assumptions upon the type of documents (e.g. historical, newspapers, etc.).

In this paper we present the first web-based system for ground truth generation (WebGT) [1]. Our system provides a user-friendly interface for quick annotation of degraded documents in general, and historical document images in particular (see Fig. 2). WebGT utilizes the web environment to bridge the gap between different platforms. Hence, researchers around the globe can use the suggested system independently of the underlying operating system on their workstations and without any special configurations or installation. It provides a convenient environment for collaborative ground truth generation, so that multiple users from different locations can be involved in this process. The data collections are managed such
that researchers can upload their own document collections and control their privacy settings. Furthermore, researchers may assign annotation tasks to different users, and approve their annotation later on. This property lays ground for collaborative work on data collections. WebGT guarantees a smooth flow of the ground truth process by avoiding the switching between pop-up dialog boxes and the annotation of ground truth.

The on-line semi-automatic strategy is one of the key innovations of the suggested system. Off-line semi-automatic systems, such as [4], enable users to correct pre-produced ground truth at a separate stage. Correcting pre-produced data was shown to be more time-consuming and error prone than starting a completely new ground truth work [2]. On-line semi-automatic systems, on the other hand, provide an instant interaction between the users and the system, namely, users create and correct ground truth on the fly. To the best knowledge of the authors, WebGT is the first on-line semi-automatic ground truth generation system. It enables users to click a considered component and automatically defines the component’s outline. Thus avoiding tedious post-processing steps, and enhancing the quality of the final result.

II. THE SYSTEM

Being a web based annotation system, users with proper permission merely need a browser to login into the system. Once a user is logged in, the system displays the collections assigned to them. Depending on user permissions, management functionalities, such as assigning a specific annotator to a group of pages, are among other options provided by the system.

The documents in the system are arranged in collections. Each collection may be accessible by different users. A collection contains manuscripts, and manuscripts may contain hundreds of pages. Each collection of documents is assigned a ground truth generation scheme by the collection manager. This scheme defines the hierarchical relationships between the document elements (such as letters, words, etc.) and specifies the resolution of the annotation (character annotation, word annotation, etc.). Fig. 3 illustrates the system creating ground truth for a text line in Arabic script.

A. The annotation process

For ease of use, a set of functionalities assist the user in the annotation process. Users can annotate the most basic elements and group elements at each stage to create higher hierarchy elements. The system is designed to minimize the usage of pop up dialogs. It enables the user to perform all the required actions using the mouse and the keyboard. This guarantees a fast and smooth navigation across the document. The system provides instant visual feedback presented over the annotated document for each action. All actions may be invoked using keyboard shortcuts.

The system has three main modes: Selection mode, Grouping mode and Annotation mode. Next, we describe each mode separately.

1) The selection Mode: Following a bottom-up approach, the process starts with selecting the basic elements of the document. Let us assume that our basic units are letters. A semi-automatic process is employed to select a single character. The user clicks on any pixel of the character using the mouse, and the system automatically captures the connected component corresponding to the pixel and its bounding box. The system runs this process in the background using a dynamic gray level threshold. The adjustment of the threshold is done by using the keyboard or by buttons located at the top of the screen.

Threshold may change across the document due to variating degradation which is a prevalent phenomenon in degraded historical documents. The system enables the user to observe the influence of the threshold adjustment upon the element selection in real time.

This semi-automatic step might produce inaccurate character selection, especially for severely degraded documents. To overcome such inaccuracies, the system enables the user to manually modify the borders of the automatically computed bounding box. Moreover, the system offers an option of fully manual element selection by the user.

Special functionalities are provided to cope with challenges such as touching components and diacritic symbols. While diacritic symbols do not have a meaning as stand-alone components, touching components may disrupt the recognition process as they are not real components of the language. To
this end, a split line can be drawn by the user to separate touching components (see Fig. 4). The generated bounding box is marked directly over the document. Similarly, the user can join two unconnected components into one by drawing a connecting line between them (see Fig. 5).

2) The grouping Mode: Adhering to the bottom-up methodology the system has a grouping mode where the user can group low elements in the element hierarchy into higher ones. The user wraps the relevant elements with a grouping rectangle using the mouse (see, e.g., Fig. 3). The ground truth generation scheme dictates which element types can be grouped together (for example grouping letters into words). Similarly, the non basic elements can be grouped further to form higher hierarchy elements, i.e. words to lines, and lines to text regions.

3) The annotation Mode: The system enables attaching annotation to basic elements according to the ground truth generation scheme. After marking the basic elements in the document, the user changes the system mode from selection to annotation. The user inserts the transcription of elements one after the other. The system superimposes the transcribed letters above the corresponding element on the screen. The annotation is stored as Unicode in the system’s database.

B. System Management

To enable an easy collaboration between different users who are involved in the ground truth generation, the system allows assigning two types of users for a collection, manager and regular user. A user can be a manager of a specific collection, while, for other collections he can be defined as a regular user. Managers are allowed to create their own private collections of documents. They can expose their documents to a set of selected users, dispatching the job of ground truth generation among them, and monitoring the progress in the annotation process. Generally, the document managing interface displays to each user either collections managed by him or documents which have been assigned to him as an annotator (see Fig. 6). The system interface enables the users to browse through these collections. Managers can add/remove collections and manuscripts.

The collection manager controls the ground truth generation scheme which represents the policy of the collection annotation. This includes setting the types of the document elements, the hierarchy among the elements, which hierarchy level will include the annotation, and whether the annotation will be restricted to one character or not. Collection managers can define a default set of metadata properties which will be attached to all the manuscripts in the considered collection (see Fig. 7 where collection settings dialog box is displayed on the screen). A regular user can add metadata properties for a specific manuscript independently, however, he cannot remove any property from the default set.

Collection Managers can export the ground truth data of documents as output files. For now the system supports two file formats, i.e., the PAGE XML format [6] used in the IMPACT project, and comma separated (CSV) files. However, the system design is flexible and it can support any XML format which will be required by users in the future.

C. Implementation

The system is built as an Ajax web application, which performs most of the processing on the client side using advanced technologies such as html5, modern java script libraries, Ajax, and secured web services. These technologies enable swift, responsive, and advanced behavior of the web application. The annotation data is continuously sent to the server during the annotation process, using a web service, and are stored in MS SQL data base. Storing the ground truth data in a database guarantees a safe annotation process which is robust against system failures, and enable fast retrieval in various views.
III. CONCLUSIONS AND FUTURE RESEARCH

We present the first web-based system for ground truth generation (WebGT). It provides a user-friendly interface for quick annotation of degraded historical document images. Usage of the web interface and the on-line semi-automatic strategy are some of the key innovations of the developed system. This tool is a cross-platform tool and can work on any machine regardless of the underlying operating system. Users are not required to deal with special configurations of their machines nor with installation issues. The instant interaction between the user and the system guarantees a reliable ground truth generation. Researchers can upload their own document collections and control their privacy settings. Furthermore, the system supports a strong collaboration between researchers of many research areas, such as historians and computer scientists, in remote locations. Researchers have the ability to assign annotation tasks to different users, and to approve their annotation later on. Our system introduces a set of novel concepts which guarantee a robust ground truth generation and a smooth human-computer interaction as well.

WebGT was tested and applied in our lab by our collaborators from the Department of Jewish Thoughts resulting in transcription of the letters aligned with the document (see Fig. 1). They report that the system is easy to use, intuitive and user friendly. Ground truth generation of one page of 300 handwritten letters in a highly damaged document took 40 minutes. The ground truth generation included defining the outline of letters, words, lines and annotating the letters.

In our future work we intend to improve the system so that it would support ground truth generation for more page layout entities (e.g. figures). Moreover, we plan to add more selection options, e.g., employing polygons instead of bounding boxes. The architecture of the existing system allows to integrate these options easily in later versions. An extended evaluation scheme will be conducted to measure different aspects of our system.

ACKNOWLEDGMENT

This research was supported in part by the DFG-Trilateral grant no. FI 1494/3-2, the Ministry of Science and Technology of Israel, the Council of Higher Education of Israel, the Lynn and William Frankel Center for Computer Sciences and by the Paul Ivanier Center for Robotics and Production Management at Ben-Gurion University, Israel. We also thank Prof. Uri Ehrlich and Uri Safrai from the Goldstein-Goren Department of Jewish Thoughts, Ben-Gurion University of the Negev, for their assistance in testing the system.

REFERENCES