IDIAP RESEARCH REPORT

MAYA CODICAL GLYPH SEGMENTATION: A CROWDSOURCING APPROACH

Gulcan Can Jean-Marc Odobez
Daniel Gatica-Perez

Idiap-RR-01-2017

JANUARY 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Maya Codical Glyph Segmentation: A Crowdsourcing Approach

Gulcan Can, Jean-Marc Odobez, Daniel Gatica-Perez

***Abstract*—This paper focuses on the curation of individual Maya glyphs from three genuine codices by the help of the crowd. More precisely, non-expert annotators are asked to segment glyph-blocks into their constituent glyph entities (based on the supervision provided by available variations of glyphs from existing expert catalogs). Compared to object recognition in natural images or medieval handwriting transcription tasks, designing an engaging task and dealing with crowd behavior is challenging in our case due to the inherently complex structure of Maya writing and an incomplete understanding of the signs and semantics in the existing catalogs. We elaborate on the evolution of the crowdsourcing task design, and discuss the choices for providing supervision during the task. We analyze the variations of similarity scores, task difficulty scores, and segmentation performance of the crowd. A unique dataset of over 9K Maya glyphs from 276 categories individually segmented from the three codices has been created and will be made publicly available thanks to this process, along with baseline methods for glyph classification using convolutional neural networks.**

***Index Terms*—crowdsourcing, Maya glyph, segmentation**

I. Introduction

Crowdsourcing is an active area in multimedia to generate labels for images and videos [24], [4], [31], [34], [37]. Recently, several large-scale databases have been curated via crowdsourcing and this allowed many advances in the multimedia and computer vision fields. Tagging images, recognizing and marking object boundaries, describing scenes or actions are several use-cases for image understanding tasks that require large-scale collaboratively-collected datasets. Similarly, advances in optical character recognition, handwriting recognition, and historical document transcription are due to the availability of large-scale datasets like MNIST [25], IAM handwriting database [27] and IAM historical document database [16], and many individual transcription projects, e.g. the Transcriptorium project [18].

Transcription tasks from handwritten or printed documents that come from different eras are studied commonly in the digital humanities literature. The fundamental step for these tasks is the generation of datasets that require digitization of the documents, transcription, and correction of uncertain situations and of human errors during transcription to finally establish the ground truth for the data. Several projects involved non-expert crowd workers in the different phases of this process, such as scanning the documents, locating the regions of interest, adding the digital entries of the data, verifying or editing other contributors' responses, etc.

In this paper, we describe the collaborative work of non-experts to build a Maya codical glyph database by locating the regions corresponding to individual glyphs within glyph-blocks. The task is defined as marking the individual glyph regions within glyph blocks given the set of variations of each glyph sign contained in these blocks, which are obtained from the existing Maya catalogs created by experts [39], [29]. This task design was possible as the annotations of the glyphs and the scanned images of the codices were previously produced by experts.

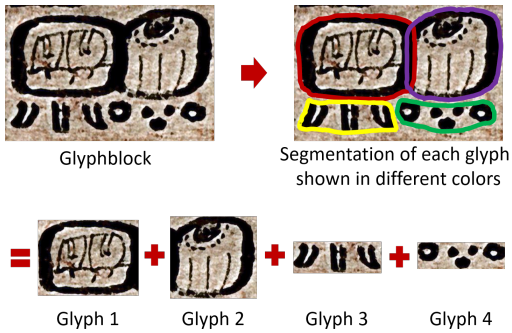


Fig. 1: Illustration of the segmentation of individual glyphs out of a glyph-block.

A main challenge for obtaining a large-scale dataset via crowdsourcing is finding and training a crowd for the specific task. Many of the large-scale digitization/transcription projects are voluntary, due to the lack of resources and vast amount of documents. An alternative approach is to leverage crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower. These two approaches differ in terms of motivation and engagement of the annotators, the number of annotators available and, in general, the amount of time needed to achieve the annotation task. With paid crowdsourcing platforms, the annotation period is generally shorter, as the crowd is gathered by the platform, and the monetary motivation is the driving force. Due to this, careful task design and annotator behavior analysis are required.

From a task perspective, glyph segmentation (illustrated in Fig. 1) is more challenging than labeling or segmenting natural images. First of all, the crowd might have never been exposed to any ancient writing system before, whereas humans interact and learn about their surroundings from an early age and have an intuition for object categories (even unseen ones) based on the similarities to already known objects. Second, the Maya language can be visually quite complex compared to other ancient writings. For instance, Egyptian hieroglyphs are usually in the form of well-separated glyphs. In Maya writings, glyph boundaries are shared between neighbors, the signs can exhibit many deformations, and some inner details

Manuscript received December 31, 2016; revised January 19, 2017.

are not always visible. Third, there are uncertainties about the categories of some signs due to severe damage, incomplete understanding of the changing shape of signs among different eras and places, and unclear semantic relationships of non-frequent signs.

The focus of this work is on producing individual glyph shape data from the three original Maya Codices (Dresden, Madrid, Paris) via online crowdsourcing. We present our design of the crowdsourcing task, investigating the effects of several features like the task definition, the use of different classic catalogs (Thompson and Macri-Vail) as glyph pattern models, and the relationship between the number of annotators, the sample complexity, and the reliability of the generated ground truth.

In summary, the contributions of this paper are five-fold:

1) design of a new crowd task;
2) assessment of the non-experts' performance for glyph recognition;
3) evaluation of closeness of the catalog samples to the Codices signs;
4) construction of a new segmented 10K glyph dataset that will be made publicly available. To our knowledge, this will be the largest public database of glyphs.
5) baseline glyph classification task, using deep learning to illustrate the challenges of the new dataset.

From our experiments, we observed that in spite of the glyph complexity, two non-expert annotations are enough in the majority of the cases to produce a consensual segmentation: For around 85% of the glyph cases, two contributors agree on the marked area more than 80%. We also observe that in the later stages of the task, as the contributors get exposed to more glyph data, the segmentation results improve. Additionally, we show that the characteristics and the similarity of the proposed glyph sign variants, and the level of the damage are the confusing points for the contributors as expected.

The rest of the paper is organized in eight sections. Section II briefly describes the Maya writing system. Section III discusses the related work on crowdsourcing and its applications in multimedia, computer vision, and digital humanities. Section IV describes the datasets used in our experiments. Section V explains the design and evolution of our crowdsourcing task. In Section VI, the details of the experimental procedure are provided. In Section VII, the annotations are analyzed with respect to key aspects in the pipeline. Section VIII presents the baseline glyph classification results obtained on the produced dataset. Finally, Section IX concludes the paper.

II. Maya Writing

The ancient Maya civilizations flourished from around 2000 BC to 1600 AD and left a great amount of cultural heritage materials. These can be found in the shape of stone monument inscriptions, folded codex pages, or ceramic items. The common ground of all these materials are the Mayan hieroglyphs, in short glyphs, written on them. The Maya writing system is visually complex, and new glyphs are discovered with almost every new archaeological site study. This brings the necessity of better digital preservation and storage systems as well as better visual and semantic analysis. Besides, the annotation of some glyphs is still open to discussion between scholars due to either visual differences or semantic analysis.

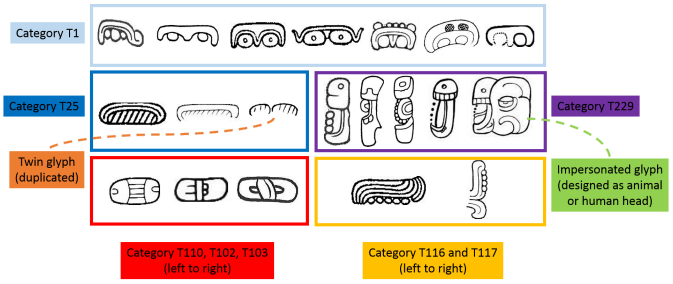


Fig. 2: Selected Maya glyph samples from several categories that illustrate the within-class variety (first two rows) and between-class similarity (last row).

Some glyphs are damaged or have many variations due to space limitations, artistic reasons, and the evolving nature of language, i.e., differences with the era and place in which glyphs were produced. Fig. 2 shows the variations of some glyphs in the top two rows. On the other hand, the boxes in the bottom row show inter-class visual differences that can be quite subtle for some categories.

For experts to decipher a glyph as a specific sign, the sign needs to have its "diagnostic part" visible or need to co-occur with another known sign to complete the semantic meaning. The expert knowledge about "diagnostic parts" and semantic co-occurrence relations has accumulated over more than a century of scholarly discussion and still continues with each new monumental site finding.

One challenging part of our study is that there are only three genuine codices today. Among these three codices, the signs are generally consistent. However, since the codices are from the post-classical era (950-1539 AD), the writing may show both simplification and variation compared to the examples found on monuments from earlier times. In terms of preservation, the folded pages of glyphs also suffered from the damage of time and external factors. Some codex pages are found partially or completely deteriorated.

A typical codex page is composed of "t'ols", which are chapter-like units composed of icons, text, and calendric signs. The text areas are composed of glyph-blocks structured as a grid. In this paper, we focus only on decomposing the text region, and more precisely, in segmenting individual glyphs out of glyph-blocks. Note that in the three codices that we study here, there is a maximum of six glyphs in a single block. This point enables to envision to have this segmentation task being achieved by non-experts with carefully-designed support.

III. Related Work

Crowdsourcing has found many applications in multimedia, computer vision, and digital humanities. Below, we briefly list several successful cases, before discussing the main challenges related to the task design, and the annotation and annotator reliability.

Large-scale Crowdsourcing Tasks in Multimedia and Computer Vision. Several widely-used benchmarks have been produced via crowdsourcing for recognition, detection, segmentation, and attribute annotation tasks. We can list as example Imagenet [35], Microsoft Common Objects in Context (MS COCO) [26], SUN scene dataset [45], SUN attribute dataset [33], or Caltech-USCD Birds-200 dataset (CUB-200-2011) [43]. Thanks to these large-scale datasets collected by the help of the crowd, more capable models were trained and advancements became possible in multimedia and vision.

Crowd workers motivated by monetary rewards (as in the case of commercial crowdsourcing platforms) as well as volunteers (in the name of citizen science) were able to generate adequate quality of content for generic object, scene, and action recognition. There has been further crowd content generation studies in sketch recognition [14] and even in specialized areas such as biomedical imaging [20], [21], [23] and astronomy (Galaxy Zoo and Zooniverse projects [17]).

Apart from content generation, crowdsourcing is also used for preprocessing of data, validation of results, and providing help for automatic algorithms. For semantic segmentation, object localization and co-segmentation, the work in [9] showed how crowd interaction helps to speed up segmentations out of bounding boxes of objects. In [11], crowdsourcing helps with clustering of the data.

Task Design. There have been studies discussing the cost-effective design of crowdsourcing tasks, and the different criteria needed to optimize cost or informativeness of the results [41], [13]. Gottlieb et. al. discuss the key elements in designing crowdtasks for satisfactory outcomes even for relatively difficult tasks [19]. They emphasize the importance of clear instructions, feedback mechanisms, and verification by qualified annotators.

The typical crowdsourcing tasks follow an "annotation-correction-verification" scheme. However, it may be challenging to apply this scheme to segmentation tasks [6]. Especially, in our case, the annotators may not be familiar with the hieroglyphic signs or their perception of the shapes may differ substantially, as the crowd has not been exposed to such visual data as often as everyday life objects in natural images. In order to guarantee satisfactory outcomes, the verification step may require an expert.

Crowdsourcing in Digital Humanities. Crowdsourcing is also adopted in Digital Humanities studies for several purposes. Digitization and transcription of historical documents with the help of the crowd is a widely-studied task. A well-known application of this task is the "re-captcha" paradigm that utilizes automated document analysis methods while keeping human intelligence in the loop [42]. Several decades of the New York Times' archives have been digitized in this way. In similar large-scale transcription tasks [10], [8], and in archaeological research on a participatory web environment [5], crowdsourcing enabled to bring valuable historical sources to the digital era for better preservation of cultural heritage as well as for further analysis.

In a preliminary work [7], we investigated the perception of glyph shape by non-experts, e.g. whether they see closed contours as a separate glyph, or how they combine visual components, assessing it in a controlled setting. The crowdworkers were asked to localize glyphs with bounding boxes in 50 glyph-blocks collected from monuments. Two scenarios were considered, either by providing the number of glyphs within a block or not. Using Amazon Mechanical Turk as platform, block-based and worker-based objective analyses were performed to assess the difficulty of glyph-block content and the performance of workers. The results suggested that a crowdsourced approach could be feasible for glyph-blocks of moderate degrees of complexity. In this paper, we significantly go beyond our first attempt, by designing an entirely new task that exploit catalog information, visual examples, and glyph variants that guide non-experts to produce arbitrary shape segmentations, and use it to segment over 10,000 individual glyphs.

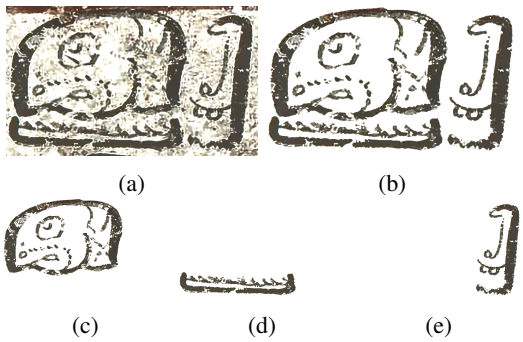


Fig. 3: The top row shows a cropped glyph-block (B1 from fifth page and second t'ol of the Dresden codex) and its cleaned image. The bottom row shows the individual glyphs in the block. These are produced by experts.

IV. Datasets

The data in our work are the glyph-blocks from three Maya Codices. To provide supervision to non-experts in our task, we utilize the glyph signs from the Thompson and Macri-Vail catalogs. The details of these datasets are given below.

A. Maya Codex Glyphs

Our sources are high-resolution images scanned from the three existing genuine Codices (Dresden [3], Madrid [1], and Paris [2]), cropped to smaller units (pages, t'ols, and glyph-blocks), and annotated with metadata by our epigrapher partners. The metadata of each glyph-block contains the name of the codex, page number, t'ol number, reading order, and relative location of the blocks in the t'ol (row and column order, i.e., A1, B2, etc.). The metadata of each glyph within each glyph-block contains its reading position in the block, its sign code from various catalogs (Thompson [39], Macri-Vail [29], Evrenov [15], and Zimmermann [47]), its phonetic value, and its damage level. The latter ranges from 0 (undecipherable) to 4 (high quality), and indicates how identifiable the glyph is according to the expert. This is not decided only based on visual degradation, but also based on the semantics and co-occurrence with neighboring glyphs.

TABLE I: The number of elements in the three codices.

	# pages	# blocks	# glyphs	# glyphs with annotation and source image
DRE	72	2924	6932	6439
MAD	100	3254	7429	6910
PAR	18	774	1620	1373
ALL	190	6952	15981	14722

Table I summarizes the number of elements available from the three Codices. Some pages of these Codices are highly-damaged. Even though there are, respectively, 76, 112, and 22 (in total, 210) pages in our database, we only list the number of pages that have at least one recognizable glyph in Table I. Similarly, we have the records of 7047 glyph-blocks in total, however only 6952 of them have at least one recognizable glyph. In total, 14722 glyphs have known catalog annotations with cropped glyph-block images. Note that the experts have not provided the individual glyph images for all these glyphs, as the segmentation of Codices into glyph-blocks and individual glyphs with high quality is quite demanding in terms of time and effort. The experts upscale and apply some preprocessing (i.e. unsharpening, and binarization) to block images with commercial tools, which requires manual handling of each block or glyph element. Furthermore, deciding annotations of glyphs for several catalogs, assigning identifiability ranking, and providing spellings are quite time-consuming. As the experts' focus is on decipherment, only a very small proportion of individual glyph segmentations has been previously produced by experts [22] (see Fig. 3). At the large scale, the experts provided only the cropped block images (as in Fig. 3a) without binarization. Therefore, we designed a crowdsourcing task for segmenting the individual glyphs out of the blocks.

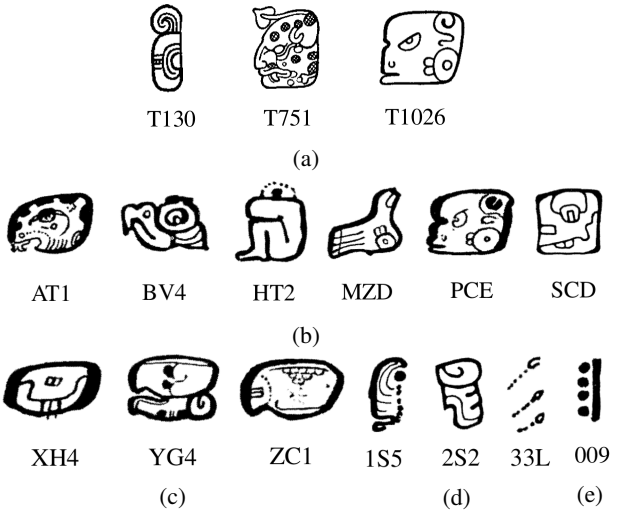


Fig. 4: (a) Examples for affix, main, and portrait Thompson categories. Last two rows illustrates examples for main Macri-Vail categories: (b) semantic (animals, body parts, and faces); (c) square (symmetric, asymmetric, and with irregular shapes); (d) elongated (with 1, 2, and 3 components); and (e) numeric categories.

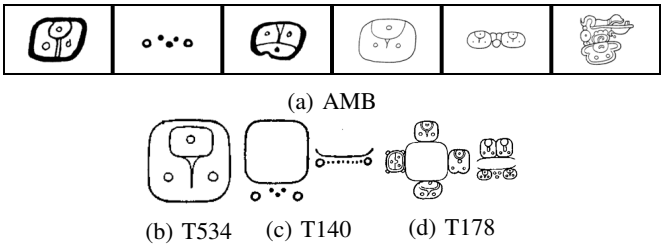


Fig. 5: (a) Variants of AMB category in Macri-Vail catalog; (b-d) occurrences of these variants in 3 different categories in the Thompson catalog.

B. Catalog Signs

The documentation of Maya writing started during the Spanish conquest of Yucatan in the XVI^{th} century. Bishop Diego de Landa's incomplete alphabet, in his book titled "Relación de las cosas de Yucatán" [12], [40], was created by asking two locals how to write Spanish characters in Maya language [44]. Later on, for several centuries, few attempts were made to understand Mayan writings. Evrenov's [15] and Thompson's [39] sign catalogs became important sources, suggesting syllabic readings rather than character correspondences of the signs. For historical reasons, Thompson's taxonomy (main and affix syllabic signs) became more influential than Evrenov's, for several decades. With the advancement of the understanding of the semantics of the signs, more modern catalogs emerged [28], [29].

The Thompson catalog has three main categories: affix, main, and portrait signs. Macri-Vail taxonomy has 13 main categories [29]. Six of them, i.e. animals, birds, body parts, hands, human faces, and supernatural faces, are grouped semantically (see Fig. 4b). There is a main category for numericals signs that are composed of dots and bars (Fig. 4e). The rest are grouped based on visual elements (square signs divided based on symmetry, e.g. Fig. 4c, and elongated signs divided based on the number of components, e.g. Fig. 4d).

Since Thompson's catalog was highly adopted for a long time and Macri-Vail's catalog has a modern taxonomy with a focus on Codices signs, we use these two resources. The fundamental difference between them is the emphasis given to visual appearance and to semantics. Thompson is known to categorize the glyphs with respect to similarity based on hand-prepared graphic cards. Macri-Vail consider co-occurrences of the signs and modern knowledge of the semantics and usage of some signs rather than visual cues only. This leads to a higher visual within-class dissimilarity of Macri-Vail signs. For instance, the variants in the AMB category (Fig. 5a) are spread over three Thompson categories (T534 main sign, T140 and T178 affix signs, see Fig. 5b-5d).

The individual glyph variants that we used in our work were obtained through manual segmentation of high-quality scanned pages of these two catalogs by our team members. As some of the numeric signs were missing in these catalogs, we manually generated them by combination of dots and lines from existing

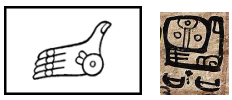


Fig. 6: An articulated hand sign (T670) from Thompson catalog and an instance of it from Paris codex.

number signs.

Utilizing these variants in a crowdsourcing task has not been previously attempted. Gathering crowd-generated assessments of the similarity of glyph variants to codex glyph samples is quite valuable in terms of eliminating one-man errors and providing finer-grained class information.

V. Crowdsourcing Task

Automatic glyph recognition starts with obtaining segmented, cleaned, and binarized glyph data. However, this is a tedious process consuming time for epigraphers. We have investigated whether the first part of this preprocessing task (glyph segmentation) can be crowdsourced. In our work, non-experts were asked to segment individual glyphs from the original glyph-block sources. Our experimental design evolved over three stages (**preliminary, small, large**). In the preliminary stage, we segmented few glyphs (27 from 10 blocks) with two different task designs. This stage helped to finalize the task design. The small stage consists of segmenting glyphs that have grountruth. This stage helped to judge which catalog was more helpful to non-experts in our task. As the large stage, we conducted the segmentation task for over 10K glyphs.

In our early experiments (preliminary- and small-scale), we utilized a subset of these glyphs (from 322 blocks) which have ground truth masks produced by experts [22]. We randomly picked 10 blocks for preliminary experiments for finalizing task design (see Section V-C). Overall, we have collected segmentations for 10949 glyphs (823 in the small-scale and 10126 in the large-scale job). These glyphs come from 4894 blocks and are distributed over 287 Macri-Vail categories.

In this section, we explain the process that led to the design of the final task. First, we describe the requirements, and present the platform used for experiments. We then discuss the early experience on the task design. We finally describe the definite version of the task.

A. Requirements

Given the annotations in the glyph-blocks (provided by epigraphy experts), and the example sign variants (taken from the catalogs), we expect the crowd to segment each individual sign in a block. As Maya glyphs can be found in articulated forms, i.e. hand signs (see Fig. 6), cropping glyph regions via bounding box may end up with inclusion of some parts from the neighbor glyphs. Therefore, for better localization, we designed the segmentation process to be done as a free-polygon rather than a bounding box.

To guide the process, we show the contributors the different variants of the sign to be segmented. As validation information, we would like to know the sign variant that the annotator chose as template to segment each glyph, and how similar the person found the chosen variant to the marked region. This can be used to verify the expert annotations and detect any outliers, where none of the provided sign variants match the block content. To account for this, we propose a "None" option along with the existing sign variants.

Another point to analyze is the perception of damage by the non-experts. Even though experts have provided a damage score for each glyph, this score shows how decipherable the glyph is, and so it is affected by the glyph co-occurrence and semantics. Non-expert perception of damage depends solely on visual appearance of the glyphs. This helps to obtain a damage score that is not affected by prior expert knowledge. It can also be used as a hint to assess the task difficulty.

The difficulty of our task is not uniform across the different glyph categories. According to the visual similarity to the variants and the damage ratio of the glyph, the task can be ambiguous and hard. To assess this, we ask the workers to provide a score for the task difficulty.

B. Platform

Terminology. We utilized the Crowdflower (CF) platform for our experiments. In the CF terminology, a *job* refers to the whole annotation process. An annotation unit is called *task*. A *page* is a set of unit tasks that a contributor needs to complete to get paid. N_t denotes the number of tasks in a page. The number of judgments per task N_j corresponds to the number of workers to annotate a single task. Workers in CF are called *contributors*. There are three levels of contributors. The level of a contributor is based on the expertise and performance in previous tasks.

To set up a job, a job owner must first define the set of data to be annotated. Then, s/he designs the task by specifying the queries that the contributors are asked to complete. The queries in the task can vary from simple text input to performing image annotations. After the task design is finalized, the job owner can curate *test questions* (TQ) to enable *quiz mode* in the job to ensure the quality of the results. Test questions are prepared by the job owner by listing acceptable answers for each query in the task. In the case where the contributor gives an answer out of the acceptable answers, the contributor fails the test question. For the image annotation query, the job owner provides a ground truth polygon over the image and sets a minimum acceptable intersection-over-union (IU) threshold. IU measure between segment S and ground truth G is defined as follows:

$$IU = \frac{|S \cap G|}{|S \cup G|} \tag{1}$$

If a contributor marks a region that overlaps with the ground truth region below the IU threshold, the contributor fails the test question, and cannot take on more tasks in the job. To activate the quiz mode, the job owner has to provide a certain percentage of the actual data as test questions. Contributors have to pass one page of the task in quiz mode before being admitted to the *work mode*, in which they work on the actual set of questions (AQ) and get paid. There is also a test question on each page in work mode. This check is effective to eliminate random responses.

The platform also provides other quality control checks. Job owners can set the minimum time to be spent on the task, the minimum accuracy that a contributor needs to achieve, and the maximum number of tasks that can be annotated by a single contributor.

After creating the answers for the test questions and fixing the job settings, the job owner launches the job, and can monitor the progress of the crowd workers.

Channels. CF has its own subscribers and is referred to as the Crowdflower-elite (CF-elite) channel. Apart from that, workers from other crowdsourcing platforms (also called channels) can also link their accounts and work on available CF jobs. This allows crowd diversity in the platform. These external platforms can be large-scale with global subscribers such as ClixSense, or can be medium- or small-scale with a focused crowd in particular countries. The choice of platforms is given to the job owner.

Platform limitations. One reason for utilizing CF is the unavailability of Amazon Mechanical Turk out of USA due to current regulations. Another reason is the readily available image annotation tool. Unfortunately, this tool does not support any validation checks. This was an important issue in our task design: The lack of controls on the number of polygons drawn in the pane or the minimum/maximum areas in pixels can be spam-prone if results are not monitored.

C. Design Experiences

We conducted 4 preliminary experiments before deciding the final task design and settings. The different settings are given in Table II, and discussed below.

Block-based design vs. glyph-based design. In the first two experiments, the initial design (shown in Fig. 7) aimed to collect *all* glyph segmentations of a glyph-block in the same task (one glyph after another in separate drawing panels). This initial design was confusing. Some workers marked all the glyph regions in the first drawing pane, instead of drawing them separately. Another source of confusion was the order of the glyphs. Learning from this, we simplified the task as *individual* glyph drawing. As a result, the average f-measure between the convex hull of a crowd-generated segmentation and the ground truth improved by more than 10% (see Table II), when moving from multi glyph annotations (75.2% and 79.5%) to the single glyph case (89.7% and 92%). More specifically, the f-measure of segment S and ground truth G is defined based on precision p and recall r as follows:

$$f = 2 * \frac{p * r}{p + r} \quad \text{as} \quad p = \frac{|S \cap G|}{|S|} \quad \text{and} \quad r = \frac{|S \cap G|}{|G|} \quad (2)$$

Number of glyph variants. We limited the number of glyph variants of each individual glyph shown to the contributors to keep them focused on the segmentation task. At first, we experimented with a maximum of three variants chosen a priori by visual clustering (12% of the signs in the Thompson catalog had more than 3 variants). After empirically verifying that increasing the number of provided variants did not hinder worker performance overall and gave them more visual cues about the possible variations, we choose to provide a maximum of six variants (if available).

Design of feedback part. In the initial design, we asked contributors about damage level as well as wrong or missing annotations. This part was often omitted by the workers. From this experience, we only kept the most direct rating factors (damage, task difficulty). We also included a text box for optional comments. Received comments included points about rotations of the glyph variants, uncertainty about the damage rating, confusion about the variant choice when several variants shared similarities with the target glyph. Based on these comments, we improved the instructions.

Crowd expertise, number of tasks per page, and payment. In the first experiment, we allowed contributors with medium- and high-level of expertise and set the payment per page as \$0.15. We hypothesized that 10 tasks per page were too many considering the payment. We observed that only medium-level contributors took the job and only 60.9% of the glyph segmentations were saved with an average f-measure of 75.2%. In the second experiment, we decreased N_t to 2, set the payment per page to \$0.30, and only allowed expert contributors (level-3). This resulted in 79.9% saved segmentation with average f-measure of 79.5%. Considering that there are three glyphs in glyph-blocks in average, we set the payment as \$0.10 for the last two glyph-based experiments to maintain payment/time ratio. Together with the simplified design and introduction of test questions, this payment and level of expertise setting brought the saved segmentation ratio very close to 100% (97.3% for the third experiment and 100% for the fourth one) with an average f-measure of around 90%.

Number of judgments. In the first experiment, we started with 10 judgment per task ($N_j = 10$). Based on the first experiment, we decided to collect less number of judgments with higher quality. Therefore, we decreased N_j to 5 in the next experiments, and improved the level of expertise and payment settings as explained above.

Crowdflower-elite channel vs. other channels. We experimented with the crowd from different channels (CF-elite channel compared to other channels) in the last two experiments. However, with the simplified individual glyph-based design, and with the level-3 contributors, we did not experience a significant difference in the segmentation scores from these separate channels (89.7% vs. 92%, see Table II).

D. Final Task

1) Overview: Based on the outcome of the preliminary experiments, we designed the final task as follows. It comprised two parts. In the first one, based on the shown variants, contributors were asked to segment (draw a tight free-hand polygon around) a similar region in the glyph-block. In the second part, contributors were asked to indicate which variant they used as template to do the segmentation, and rate how similar it was to the segmented region, how damaged the glyph region is, and how easy it was to complete the task. These ratings are designed on a scale of 5.

2) Training: We provided a detailed description of the tasks, a how-to Youtube video, and positive/negative example

TABLE II: Preliminary phase segmentation results using variants of Thompson catalog.

Exp.	Catalog Variants	Block-based or glyph-based?	# Judgments per task (N_j)	# Tasks in a page (N_t)	Payment per page ($)	Min level of contributors	Allowed Channels	Average f-measure (%)
1	T	Block-based	10	10	0.15	Medium	All	75.2
2	T	Block-based	5	2	0.30	High	All	79.5
3	T	Glyph-based	5	2	0.10	High	All except CF-elite	89.7
4	T	Glyph-based	5	2	0.10	High	CF-elite	92.0

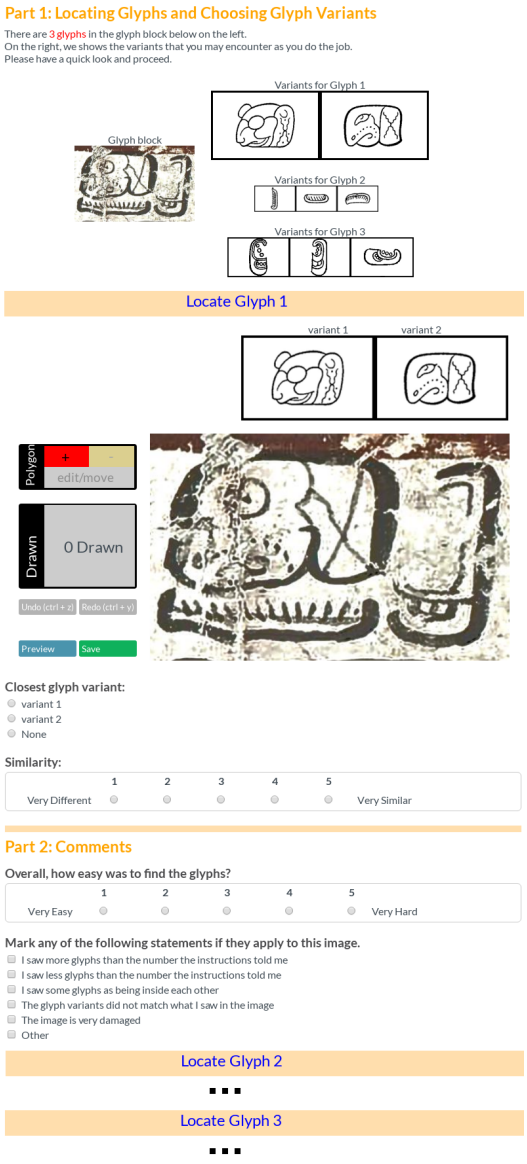


Fig. 7: Initial block-based task design (illustrating only the first glyph in the block for sake of brevity).

segmentation images, example of damage levels, and explained that segmentation quality would be checked.

3) Drawing: We used the image annotation instance tool in Crowdflower for free polygon drawing over the glyph-block images. This tool allows correction, and multiple polygons, which is useful for glyph repetition cases.

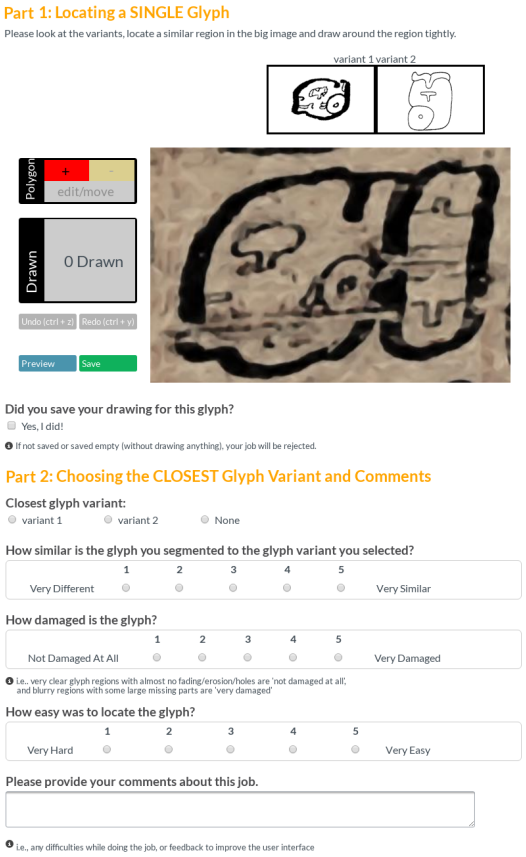


Fig. 8: Final task design.

4) Evaluation: We selected the quiz mode for the jobs: we provided tasks with known answers (ground truth polygons) and a quality threshold on intersection-over-union (IU) measure (see Section V-B) to filter out spammers and ensure quality.

VI. Experimental Protocol

Given the decisions made on the interface during the preliminary stage, we first conduct the small-scale stage over the glyphs which have ground truth, and then we run the large-scale stage. This section explains the settings of these two stages briefly.

A. Small-scale stage

In this stage, we run two experiments whose parameters are summarized in Table III. For the 823 individual glyphs (322 blocks) that have expert ground truth masks, we set up the task

TABLE III: Experimental settings for the small-scale and large-scale stages.

Exp.	Cat. Var.	# Judg. per task (N_j)	# Tasks per page (N_t)	Pay. per page ($)	# pages	IU th.
S-1	T	5	2	0.10	338	0.7
S-2	MV	5	2	0.10	344	0.7
L-1	MV	2	4	0.16	1670	0.7
L-2	MV	2	4	0.16	1732	0.8

with 1) Thompson (T), and 2) Macri-Vail (MV) references of the glyphs. In other words, in the task design, we display the glyph variants either from the Thompson or the Macri-Vail catalog.

In both cases, the number of judgments N_j is set to 5. The minimum acceptable IU score is set to 0.7. The minimum time to be spent on a page is set as 30 seconds. The maximum number of judgments by a single contributor is set as 12. As a result, a single contributor annotated 5 glyphs from the actual target set and also answered 7 test questions.

B. Large-scale stage

In this stage, we define the job for all annotated glyphs for which no expert segmentation is available. To reduce the annotation cost and having confirmed that in general most of the glyphs had a high segmentation consensus (see small-scale stage analysis in Section VII-A), we decided to collect only two judgments per glyph, and collect more only if disagreement was detected. In order to reduce the latter, we decided to exclude the following glyphs from the annotation:

- too damaged glyphs according to the damage scores from the expert and visual post-inspection of a team member,
- repetition cases (multiple instances of the same glyph in the block),
- infix cases (two separate glyphs merged by modern decipherment for semantic reasons).

As a result, we obtained 10126 glyphs (out of 14722 glyphs from the available segmented glyph-block images).

For this stage, we only relied on the Macri-Vail catalog which is a more modern resource in epigraphy.

We set the minimum IU threshold to 0.7 for the first half of the glyphs (5K glyphs) and 0.8 for the rest. This threshold ensures that the contributors do a good job on the test questions (their segmentation matches with the provided ground truth), and presumably on the actual questions so that high consensus on the collected segmentations for each glyph can be obtained. We observed that we need contributors with higher performance, as we depend on the segmentations coming from only two contributors per glyph in this setting. That is why we increased the min IU threshold for the second half of the glyphs. The minimum time spent on the task was set as 30 seconds. The maximum number of judgments by a single contributor was set as 48.

C. Segmentation Evaluation Procedure

Evaluation is performed by comparing the ground truth of the glyphs with the crowd segmentations for the small-scale stage. This is detailed in Section VII-A. For the large-scale stage, we compare the segmentations of the contributors against each other. We also checked the problematic cases in which the f-measure agreement is less than 0.8 among contributors as an internal task in Crowdflower platform.

VII. Annotation Analysis

In this section, the crowd annotations for small-scale and large-scale stages are presented in terms of the analysis of ratings and segmentations.

A. Small-Scale Stage

As described in Section VI-A, we conducted two experiments in small-scale stage: 1) with Thompson (T), and 2) with Macri-Vail (MV) references of the glyphs. We analyze the annotations from these experiments in four aspects: variant selection, damage rating, segmentation analysis, and sensitivity to the number of annotators.

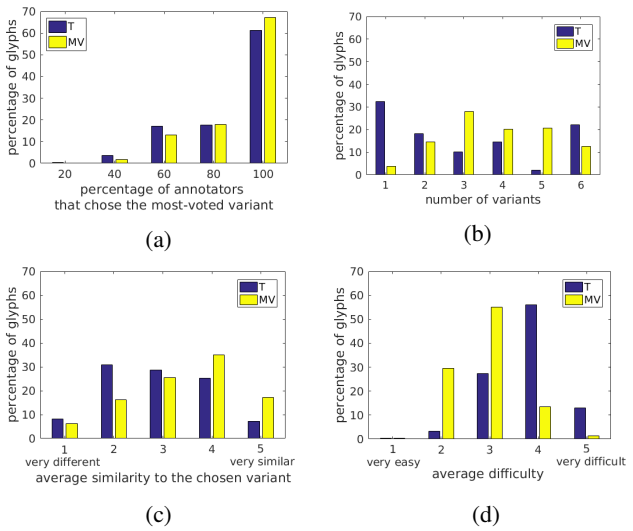
1) Variant Selection: We compare the agreement for the variant selection in the two experiments. First, note that the MV catalog contains the glyph variants from both codices and monuments, whereas the variants in the Thompson catalog come only from monuments. Typically, monumental glyphs have more details and are visually more complex than codical glyphs. In this sense, the variants from the Thompson catalog are in general more different from the codices glyphs than the MV variants.

The final variant for each glyph is selected by majority voting among the contributors' responses. Fig. 9a shows the percentage of contributors that selected the most-voted variant for the experiments with the Thompson (blue) and Macri-Vail (yellow) variants. We observe that all of the contributors agreed on a variant for 67.22% of the glyphs when the MV variants (yellow) were shown (61.22% for the T case).

Fig. 9b shows the histogram of the number of variants for the annotated glyph categories. The median values are 2 and 4 for T (blue) and MV (yellow) variants, respectively. Even though there were, in general, more number of variants available, full agreement was higher for the MV case.

A related result is illustrated in Fig. 9c. The contributors gave higher ratings of visual similarity of the chosen variant to their marked glyph region to MV variants. The mean of the average similarity ratings obtained with T variants is 2.46 (in the scale of 5), whereas this value is 2.98 with MV variants case.

Moreover, the contributors found the task harder in the case of T variants (Fig. 9d). The mean of the average difficulty ratings obtained with T variants is 3.37 (in the scale of 5), whereas this value is 2.44 with MV variants case. We also applied Kolmogorov-Smirnov non-parametric hypothesis testing [30] to compare cumulative distributions of difficulty and similarity ratings obtained with the T and MV variants. For both tests, the null hypotheses, that the average T rating



(a) (b) (c) (d)

Fig. 9: The distributions of average ratings in the small-scale stage with Thompson (blue) and Macri-Vail variants (yellow).

samples and MV rating samples come from the same distribution, are rejected at 0.01 significance level with p values of $1.2883\text{x}10^{-118}$ and $2.9343\text{x}10^{-16}$ respectively.

Overall, we observed that MV-variant tasks are rated easier, and reach higher consensus rates for the chosen variant than the T-variant cases.

2) Damage Rating: The average damage ratings (range of 1 to 5) from the crowd and the damage rating assigned by the experts are considerably different. For the experts, more than 90% of the glyphs in this set were easily recognizable (5 in the range of 1 to 5). However, the damage perception of the non-experts was focused around the middle of the scale. For around 64% of the glyphs, the contributors selected "moderate-damage" (3 in the range of 1 to 5) for both T and MV cases. This can be interpreted as the raw block crops are visually noisy in most of the cases, even though for the experts the glyphs are in good conditions to be identified.

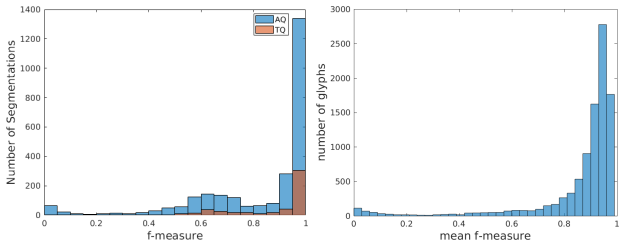
3) Segmentation Analysis: For each glyph, an aggregated mask is generated from the crowd segmentation masks such that at least half of the contributors (i.e. at least 3) marked an image point as belonging to the glyph region as illustrated in Fig. 10.

The evaluation is performed by comparing (1) the aggregated segment against the binary ground truth (S vs. GT), (2) the aggregated segment against the convex hull of the binary ground truth (S vs. GT-CH), and (3) the convex hull of the aggregated segment against the convex hull of the ground truth (S-CH vs. GT-CH). Results are shown in Table IV. We observed that most of the contributors mark the glyph regions without going into fine contour details, as it can be quite time-consuming. This is acceptable, as the main interest is in the regions with the target glyph rather than very detailed markings. Therefore, we decided to use convex hulls for further evaluation in Figs 11-12b.

Table IV summarizes the comparative segmentation performance with the help of the variants from the two catalogs. It is observed that the MV variants helped to bring out

(a) (b) (c)

Fig. 10: a) The convex hull of the ground truth (red), b) the gray-scale image of the aggregated segmentations, and c) the final aggregated segmentation for the glyph in Fig. 3e.



(a) (b)

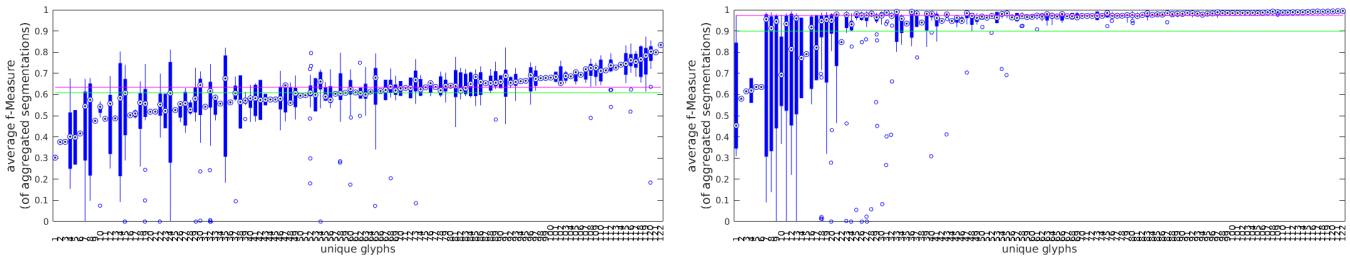
Fig. 11: a) The f-measure distributions of the segmentations (against ground truth) in actual question set (AQ, blue) and test question set (TQ, orange) with the MV variants in the small-scale experiment. b) The mean f-measure agreements for the glyphs in large-scale experiment.

TABLE IV: Average f-measure values of aggregated segmentations obtained with Thompson and Macri-Vail variants in small-scale stage.

Catalog Variants	Set	S vs. GT (%)	S vs. GT-CH (%)	S-CH vs. GT-CH (%)
T	TQ	65.7	94.5	96.6
MV	TQ	65.5	95.1	97.3
T	AQ	59.1	85.1	87.5
MV	AQ	59.9	86.4	88.6
T	All	60.2	86.6	89.0
MV	All	60.8	87.7	89.9

marginally better aggregated segmentations. The table also reports the mean scores when we consider the glyphs used as test questions (TQ) and actual questions (AQ) as separate sets. The f-measure distributions of TQ and AQ sets in the MV variants cases are plotted in Fig. 11 (the T variants case has similar distributions). We observe that the majority of the glyphs are well-segmented. As we manually chose the test questions to be relatively easy to annotate, we observe a relatively higher mean f-measure for TQ sets than for AQ sets.

Fig. 12 illustrates the boxplots of the sorted average f-score values of 122 non-numerical MV classes (left for S vs. GT, and right for S-CH vs. GT-CH comparison). While most of the classes are well-segmented, few of them have low average f-measure (5 classes have an average f-measure less than 40%). We observe that these classes are visually more complex and composed of several parts that might not have been marked by the contributors. When using the convex hull comparison, only first ten classes have an average f-score less than 70%.



(a) S vs. GT (b) S-CH vs. GT-CH

Fig. 12: Sorted average f-measure of aggregated segmentations for the unique glyph categories in the small-scale stage.

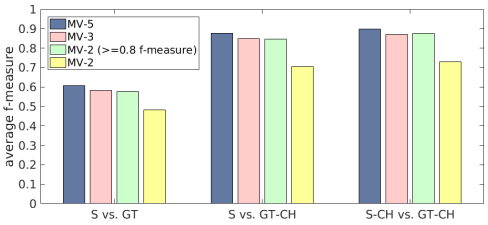


Fig. 13: Mean f-measure values of the aggregated masks obtained using 5 (blue), 3 (pink), 2 (yellow) segmentations, and 2 segmentations that have at least 0.8 f-measure agreement (green) per glyph with MV variants.

4) Sensitivity to The Number of Annotators: We simulated the performance for the case of fewer annotators. Fig. 13 shows the average f-measure values for the aggregated masks with different number of segmentations. We aggregated a maximum of 10 combinations of randomly selected segmentations, and took the mean f-score of these aggregated masks for each glyph. Obtaining aggregated masks with 3 segmentations (MV-3) rather than 5 (MV-5) resulted in a marginal decrease in the average f-score (blue to pink bars).

Furthermore, we analyzed the intersection of two segmentations either for the randomly selected ones (MV-2 yellow bars) or in the case of above 0.8 f-measure agreement (MV-2 green bars). In the latter case, we obtained very similar average f-score results to the ones with 3-segmentations. Besides, the standard deviation of the f-measures obtained with randomly sampled 2-annotations are below 10% and are usually acceptable. These observations motivated us to perform the large-scale stage with two annotations per glyph.

5) Conclusion: 368 and 397 unique contributors participated to the small-scale stage for the T-variant and MV-variant cases respectively. The corresponding average number of glyph annotations per contributor were 7.3 and 8.9 (median, min and max values were 5, 1, and 24 for the T-variant case, and 6, 1, and 29 for the MV-variant case respectively). This evaluation shows that the defined task is simple enough for a non-expert crowd to produce satisfactory results. Even though the contributors may get confused and segment parts of other glyphs which look more similar to the variants than the target region, overall the performance of the contributors was encouraging to proceed with the large-scale stage.

B. Large-Scale Stage

Here, we analyze the results observed during large-scale stage. We obtained 21907 annotations containing 20982 saved segmentations.

1) Glyph Variant Selection: Fig. 14a shows that the first variant was chosen in 73.2% of the annotations. This is not surprising as usually the two first variants in the Macri-Vail catalog are instances directly taken from the codices, and the others are line drawings of generally more complex monumental glyphs taken from the Macri-Looper catalog [28]. In 7.7% of these annotations, "none of the variants" option was chosen.

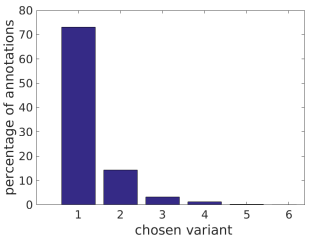
Moreover, for 23.2% of the annotations, the contributors found that the chosen variant looked different or very different than the glyph they had segmented. On the other hand, only 10.5% of the annotations are marked as "very similar." This has to be investigated further, but the reason behind it may be the tendency of the crowd to be conservative or unsure about the visual similarity scale, or indeed due to the visual differences of the glyph regions and the variants.

2) Task Difficulty and Glyph Damage: For the damage ratings, we notice that the general tendency of the contributors (41.9% of the annotations) is to give average score. However, we remark that there are still cases marked as "damaged" or "very damaged" by the non-expert crowd (30.6%), even though we provided glyph cases that are identifiable and in good condition according to the experts. We believe that the crowd gives relative ratings in the full-scale according to the examples they have previously seen.

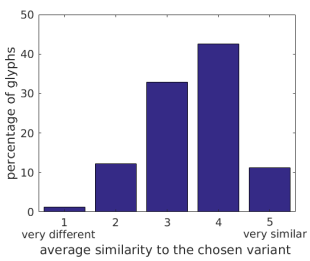
In terms of task difficulty, only 16.9% of the annotations have "hard" or "very hard' ratings. This is a positive feedback from the crowd about the perception of the task complexity.

3) Segmentation Analysis: Fig. 11b shows overall f-measure agreement distribution for the large-scale set.

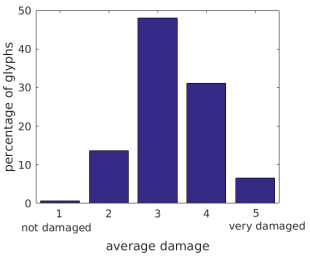
Verification. For the cases with more than 0.8 f-measure agreement, we have cropped the bounding boxes of the segmentations and scrolled through them to spot problematic cases for each sign category. There were much less problems in such cases (318 out of 8229 glyphs) in which both the contributors marked another region as the glyph area or are confused due to category annotation problems. Among these cases, we re-assigned the segmentations that correspond to neighboring glyphs, and used them while obtaining the aggregated mask. We, visually, checked the segmentations of



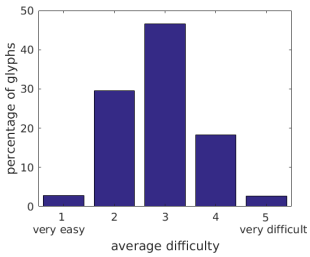
(a) The percentage of the most voted variant.



(b) Similarity ratings.



(c) Damage ratings.



(d) Difficulty ratings.

Fig. 14: Distributions of the ratings in the large-scale stage.

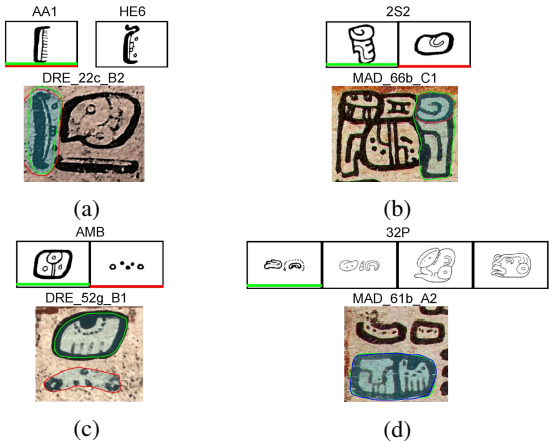


Fig. 15: Confused segmentations from the large-scale stage due to (a) similar glyphs in the block, and damaged instances, (b-c) visually-confusing variants, (d) visually dissimilar glyph instances.

1921 glyphs that have less than 0.8 f-measure agreement, and marked the segmentation that matches the target region.

Minimum IU Threshold. As described in Section VI-B, for the first half of the glyphs in the large-scale stage, the minimum intersection-over-union measure between the annotator's segmentation and the indicated segmentation of the test questions was set to 0.7. This threshold is increased to 0.8 for the rest of the glyphs. With this more strict threshold, we observed a 3.8% increase in average median f-measure agreement (from 90.2% to 94.0%) and a 5.7% increase in average mean f-measure agreement (from 82.1% to 87.7%).

Challenging Cases. Clearly, the difficulty of our task is not uniform across the glyph instances. Fig. 15 illustrates some of the cases with high disagreement between segmentations. The main reasons for high disagreement are:

- Glyph complexity: Glyphs with a large convex area are easier to segment than concave and discontinuous glyphs, i.e. with many separate parts. For instance, in Fig. 15c, one contributor selected a concave large glyph (green) somehow resembling the first variant instead of the red target region.
- Confusion from the variants: Some variants are a subset or superset of others (i.e., 2S2), as shown in Fig. 15b.
- Dissimilarity between the target region and the variants: We identify three subcases.
  - Target sample not covered by catalog variants. We noticed that some glyph instances from eight MV classes do not look similar to any of the available variants. For instance, in Fig. 15d, the target region is missed by all contributors and the neighbor glyphs were marked instead.
  - Partial dissimilarity of the glyph. Some glyph instances may exhibit different partial elements from the category variants (Fig. 15b) and this may confuse non-experts.
  - Wrong class annotation. Considering the tedious process of labeling a glyph with the codes from several catalogs, manual mislabeling is inevitable. We were able to identify few such cases.
- Mismatch of the damage rating between expert and non-expert views. This may be explained in terms of recognition of the glyph within context vs. visual completeness. For instance, in Fig. 15a, none of the contributors marked the target region, as the target region is either damaged or lacks partial details.
- Similarity of the other glyphs in the block. The red segmentation in Fig. 15a exemplifies this case, even though the target glyph belongs to class AA1, not HE6, the outline of the neighboring glyph is quite similar to the target region, and the visual difference is subtle.

4) Conclusion.: 328 unique contributors participated to the large-scale stage. The average number of glyph annotations per contributor was 66.8 (median, min and max values were 33, 2, and 432 respectively). This stage illustrates the feasibility of obtaining satisfactory outcomes even in the case of two non-experts and with minimal manual verification. Overall, we obtained valid segments for 9175 glyphs (together with the ones from the small-scale stage) that are spread over 276 MV categories. We used the aggregated valid segments in the classification task that is described in the next section.

VIII. Baseline Classification Experiments

We now illustrate how our dataset can be used in glyph classification using standard methods.

Razavian et. al. report that the penultimate fully-connected layer activations from a pre-trained convolutional neural network (CNN) are competitive with the problem-specific state-of-the-art approaches in several computer vision tasks [36]. Furthermore, the mid-layer activations have been shown to be more generic than the last-layer activations in a transfer learning setting [46]. Therefore, to obtain a baseline performance for the new glyph dataset, we conducted classification experiments on the mid-layer activations from a pre-trained deep network.

A. Data Preparation

Note that our intent is in defining and illustrating a baseline method that highlights challenges and possible classification tasks. For assesing the difficulty of our dataset, we experimented with different number of classes. We considered the glyphs with at least one valid segmentation. We have 10 classes with more than 200 such glyphs, whereas 47 classes have just one such glyph. To handle the data imbalance problem, in each experiment, we randomly picked an equal number of glyphs from each class, and we repeated this 5-times. Then, we divided each set of glyphs to training (60%), validation (20%), and test sets (20%). We report the average accuracies among 5-folds in Section VIII-C.

For each glyph, to obtain a square crop centered on the aggregated binary mask, we applied the following steps.

- Dilation. We buffered the aggregated mask via dilation in case of segmentation not covering all boundary pixels. We set the dilation amount dynamically as $1/32$ of the long edge size of the bounding box.

- Color filling. We sampled 3 red-green-blue (RGB) colors from background areas of the codices. Additionally, we computed a dynamic RGB value from each block image as $0.65 * threshold_{Otsu}$ [32]. In the need of padding, we filled the areas with these RGB values. Note that this quadruples the number of samples per class.

- Padding. For convenience during convolution, we applied padding around all the edges for $1/6$ of the long edge size of the buffered aggregated mask. Then we padded the short edge to make the final crop square-sized.

- Scaling. We scaled all processed square crops to 128x128 pixels. Note that the testing image size may differ from the training image size in VGG-net, since testing images are forwarded through only convolutional layers.

B. Methodology

We forward the preprocessed glyph segmentations in the pre-trained VGG-16 network [38] and extract conv5 (last convolutional layer) features. We trained a fully-connected shallow network (two layers) from these features to perform classification.

C. Results

Table V shows the average accuracies among 5-fold experiments with different number of classes. As the number of classes increases and the number of samples per class decreases, the classification problem becomes more challenging. With 200 glyphs per class in the 10-class experiment, we obtained 87% average accuracy. For the 150-class case, we obtained 20.9% accuracy (random baseline would be 0.66%). These results both show the complexity of the data set and encourage further transfer learning experiments with deep features even in the case of small amount of target data.

TABLE V: Average accuracies for classification experiments.

	# of classes (# of glyphs per class)					
Avg. Acc.	**10 (200)**	**25 (100)**	**30 (80)**	**50 (48)**	**100 (20)**	**150 (5)**
Train	99.9	97.8	95.8	81.5	41.7	52.4
Val.	86.2	80.3	77.0	65.4	38.6	21.8
Test	87.0	79.4	76.8	62.4	36.2	20.9

IX. Conclusion

In this work, we achieved the segmentation of Maya glyphs from three genuine codices (Dresden, Madrid, and Paris) with the help of the crowd. The main conclusions are the following:

- **Task design.** As the data target does not come from everyday-objects, guiding non-experts is essential to obtain a satisfactory outcome. From our experience with the task design in the preliminary stage, we observed that simpler and focused task design (to segment individual glyphs rather than all glyphs in a block) and clear instructions are indispensible.
- **Catalog choice.** From the small-scale stage, we concluded that the variants from the MV catalog matched a higher percentage of the glyph instances compared to the variants from the T catalog. This enabled the non-experts to reach a higher consensus on the "closest-looking" variant, and obtain higher average f-measure. Furthermore, we observed that the crowd found the task easier with MV variants. These results were expected as monumental glyphs were the main source of Thompson catalog variants. Even though Thompson catalog has more number of variants per glyph, they fail to span the codical version of the glyphs, which became a strong point of MV catalog in our work.
- **Non-expert behavior analysis.** We pointed out the main challenges that the crowd faced during the task, such as visual within-class dissimilarities or between-class similarities, and effect of damage. These challenges affect the segmentation outcome. However, they are inherent from the nature of the data. That is why our work needed a careful task design, and multi-stage analysis (preliminary, small- and large-scale).
- **Maya codical glyph corpus.** We gathered over 9K individual glyphs corpus from the three genuine Maya codices along with the corresponding metadata, such as similarity rating of the instances to the MV variants. The dataset will be made publicly available.
- **Baseline classification.** We presented baseline results for classification tasks on the new dataset. These results illustrate that the new dataset is challenging, and yet the transfer learning methods with deep neural networks are promising to explore even on this challenging dataset.

Acknowledgment

This work was funded by the SNSF MAAYA project. We thank our partners Carlos Pallán Gayol (University of Bonn), Guido Krempel (University of Bonn), Jacub Spotak (Comenius University in Bratislava) for help with the glyph block data preparation and glyph annotation, and Rui Hu (Idiap) for discussions.

REFERENCES

[1] "Madrid codex," http://www.famsi.org/mayawriting/codices/madrid.html.

[2] "Paris Codex," http://gallica.bnf.fr/ark:/12148/btv1b8446947j/f1.zoom.r=Codex

[3] "Saxon State and University Library Dresden (SLUB) library," http://digital.slub-dresden.de/werkansicht/dlf/2967/1/.

[4] J. I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, Jan 2013.

[5] C. Bonacchi, A. Bevan, D. Pett, A. Keinan-Schoonbaert, R. Sparks, J. Wexler, and N. Wilkin, "Crowd-sourced archaeological research: The micropasts project," *Archaeology International*, vol. 17, 2014.

[6] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *ECCV*. Springer, 2010, pp. 438–451.

[7] G. Can, J.-M. Odobez, and D. Gatica-Perez, "Is that a jaguar?: Segmenting ancient maya glyphs via crowdsourcing," in *International Workshop on Crowdsourcing for Multimedia*. ACM, 2014, pp. 37–40.

[8] L. Carletti, G. Giannachi, D. Price, and D. McAuley, "Digital humanities and crowdsourcing: An exploration," in *Museum and the Web*, 2013, pp. 223–236.

[9] A. Carlier, V. Charvillat, A. Salvador, X. Giro-i Nieto, and O. Marques, "Click'n'cut: crowdsourced interactive segmentation with object candidates," in *International Workshop on Crowdsourcing for Multimedia*. ACM, 2014, pp. 53–56.

[10] T. Causer and M. Terras, ""many hands make light work. many hands together make merry work": Transcribe bentham and crowdsourcing manuscript collections," M. Ridge, Ed. Ashgate Surey, 2014, pp. 57–88.

[11] Q. Chen, G. Wang, and C. L. Tan, "Web image organization and object discovery by actively creating visual clusters through crowdsourcing," in *ICTAI*, vol. 1. IEEE, 2012, pp. 419–427.

[12] C. E. B. de Bourbourg, *Relation des choses de Yucatan de Diego de Landa*. Durand, 1864.

[13] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei, "Scalable multi-label annotation," in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3099–3102.

[14] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 44:1–44:10, jul 2012.

[15] E. Evrenov, Y. Kosarev, and B. Ustinov, *The Application of Electronic Computers in Research of the Ancient Maya Writing*. USSR, Novosibirsk, 1961.

[16] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Ground truth creation for handwriting recognition in historical documents," in *IAPR International Workshop on Document Analysis Systems*. ACM, 2010, pp. 3–10.

[17] L. Fortson, K. Masters, and R. Nichol, "Galaxy zoo," *Advances in machine learning and data mining for astronomy*, vol. 2012, pp. 213–236, 2012.

[18] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. A. Sánchez, A. H. Toselli, and E. Vidal, "Ground-truth production in the transcriptorium project," in *IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 237–241.

[19] L. Gottlieb, G. Friedland, J. Choi, P. Kelm, and T. Sikora, "Creating experts from the crowd: Techniques for finding workers for difficult tasks," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 2075–2079, Nov 2014.

[20] D. Gurari, D. Theriault, M. Sameki, and M. Betke, "How to use level set methods to accurately find boundaries of cells in biomedical images? evaluation of six methods paired with automated and crowdsourced initial contours." in *MICCAI: Interactive Medical Image Computation (IMIC) Workshop*, 2014, p. 9.

[21] D. Gurari, D. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, C. Zhang, J. Y. Wong *et al.*, "How to collect segmentations for biomedical images? a benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," in *Winter Conf. on Applications of Computer Vision*. IEEE, 2015, pp. 1169–1176.

[22] R. Hu, G. Can, C. Pallan Gayol, G. Krempel, J. Spotak, G. Vail, S. Marchand-Maillet, J.-M. Odobez, and D. Gatica-Perez, "Multimedia analysis and access of ancient maya epigraphy," *Signal Processing Magazine*, vol. 32, no. 4, pp. 75–84, Jul. 2015.

[23] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, and A. H. Beck, "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd," in *Pacific Symposium on Biocomputing*. NIH, 2015, p. 294.

[24] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. Jones, "The community and the crowd: Multimedia benchmark dataset development," *IEEE MultiMedia*, vol. 19, no. 3, pp. 15–23, July 2012.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[27] M. Liwicki and H. Bunke, "Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard," in *ICDAR*. IEEE, 2005, pp. 956–961.

[28] M. J. Macri and M. G. Looper, *The New Catalog of Maya Hieroglyphs: The Classic Period Inscriptions*. University of Oklahoma Press, 2003, vol. 1.

[29] M. J. Macri and G. Vail, *The New Catalog of Maya Hieroglyphs, vol. 2: The Codical Texts*. University of Oklahoma Press, 2008.

[30] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[31] L. S. Nguyen and D. Gatica-Perez, "Hirability in the wild: Analysis of online conversational video resumes," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1422–1437, July 2016.

[32] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[33] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*. IEEE, 2012, pp. 2751–2758.

[34] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1231–1243, Oct 2013.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[36] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *CVPR Workshops*, June 2014.

[37] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, July 2016.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[39] J. E. S. Thompson and G. E. Stuart, *A Catalog of Maya Hieroglyphs*. University of Oklahoma Press, 1962.

[40] A. M. Tozzer, *Landa's Relacion de las Cosas de Yucatan: a translation*. Peabody Museum of American Archaeology and Ethnology, Harvard University, 1941.

[41] S. Vijayanarasimhan and K. Grauman, "Cost-sensitive active visual category learning," *IJCV*, vol. 91, no. 1, pp. 24–44, 2011.

[42] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[44] Wikipedia, "Diego de Landa — Wikipedia, the free encyclopedia," 2016, [accessed 10-November-2016].

[45] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*. IEEE, 2010, pp. 3485–3492.

[46] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in NIPS*, 2014, pp. 3320–3328.

[47] G. Zimmerman, "Die hieroglyphen der maya-handschriften, cram," 1956.

Gulcan Can is a PhD. Candidate at Idiap Research Institute and EPFL in Switzerland. Email: gcan@idiap.ch

Jean-Marc Odobez is the Head of the Perception and Activity Understanding group at Idiap, and Maitre d'Enseignement et de Recherche at EPFL, Switzerland. Email: odobez@idiap.ch

Daniel Gatica-Perez is the Head of the Social Computing Group at Idiap and Professor Titulaire at EPFL, Switzerland. Email: gatica@idiap.ch