MAYA CODICAL GLYPH SEGMENTATION: A CROWDSOURCING APPROACH

Gulcan Can   Jean-Marc Odobez

Daniel Gatica-Perez

Version of JANUARY 29, 2018

Maya Codical Glyph Segmentation: A Crowdsourcing Approach

Gulcan Can, Jean-Marc Odobez, *Member, IEEE,* Daniel Gatica-Perez, *Member, IEEE*

Abstract—This paper focuses on the crowd-annotation of an ancient Maya glyph dataset derived from the three ancient codices that survived up to date. More precisely, non-expert annotators are asked to segment glyph-blocks into their constituent glyph entities. As a means of supervision, available glyph variants are provided to the annotators during the crowdsourcing task. Compared to object recognition in natural images or handwriting transcription tasks, designing an engaging task and dealing with crowd behavior is challenging in our case. This challenge originates from the inherent complexity 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 distributions of similarity and task difficulty scores, and the segmentation performance of the crowd. A unique dataset of over 9000 Maya glyphs from 291 categories individually segmented from the three codices was created and will be made publicly available thanks to this process. This dataset lends itself to automatic glyph classification tasks. We provide baseline methods for glyph classification using traditional shape descriptors and convolutional neural networks.

Index Terms—crowdsourcing, Maya glyph, classification

I. INTRODUCTION

Crowdsourcing is an active area in multimedia to generate labels for images and videos [30], [4], [37], [42], [45]. Tagging images, marking object boundaries, and describing scenes or actions are use-cases for image understanding tasks that require large-scale, collaboratively-collected datasets, e.g. Imagenet [43] and MS COCO [32]. Similarly, optical character recognition and historical document transcription have advanced thanks to the availability of large-scale datasets like MNIST [31], IAM [33], [16], and many individual transcription projects [19].

In Digital Humanities, dataset generation is a fundamental step for document analysis tasks. Dataset generation requires digitization, transcription, and correction of uncertain situations and of human errors during transcription. Several projects have involved non-expert crowd workers in the different phases of this process, such as scanning documents, locating regions of interest, adding digital entries, verifying or editing other contributors' responses, etc.

In this paper, to study automatic algorithms to analyze Maya glyph shapes, we aim to build a *Maya individual codical glyph* database from the remaining codex resources. In this context, we describe the collaborative work of non-experts by locating

G. Can, J.M. Odobez, and D. Gatica-Perez are with Idiap Research Institute, and the School of Electrical Engineering of the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. e-mail: gcan@idiap.ch, odobez@idiap.ch, and gatica@idiap.ch.
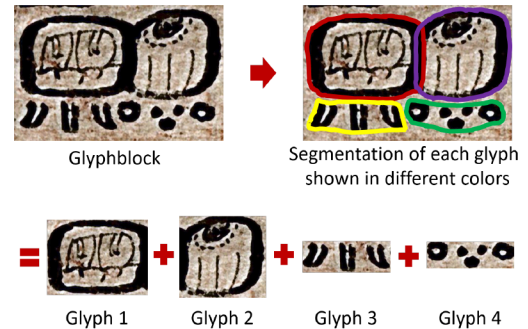

Fig. 1: Illustration of the segmentation of individual glyphs out of a glyph-block.

the regions corresponding to individual glyphs within glyph-blocks (see Fig. 1). The task is defined as marking individual glyph regions within glyph-blocks given the set of variations of each glyph sign contained in these blocks, which are obtained from existing Maya catalogs created by experts [47], [35]. This task design was possible as the textual annotations of the glyphs and the scanned images of the codices were previously produced by experts.

Crowd engagement is a challenge while curating large-scale datasets. Many 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. Therefore, 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 due to the following factors:

• **Unfamiliarity.** The participating crowd might have never seen an ancient writing system before, whereas humans interact with 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.

• **Visual Complexity.** The Maya language can be visually 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 are not always visible.

• **Uncertainty.** There are uncertainties about the categories of some signs due to severe damage, incomplete understanding of the changing shape of signs across 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 the crowdsourcing task design, 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 relations between the number of annotators, the sample complexity, and the reliability of the generated ground truth.

The main goal of generating this glyph dataset is to enable robust shape representation learning for automatic recognition tasks. Utilizing such an automatic classification or retrieval tool with reasonable accuracy, experts could identify the category of new glyph samples faster than manually going through catalogs. Furthermore, such shape representations can be used as a quantitative similarity measure. In this context, we map glyph samples into lower dimensional spaces (2-D) based on their shape representations. Such kind of mapping tools could help experts during catalog construction. This kind of tools might also facilitate discussions among scholars as part of the categorization of non-frequent glyph samples.

The contributions of this paper are three-fold:

1) *Glyph segmentation crowdsourcing:* Novel task accounting for fine-grain mapping of catalog variants to codex samples, and multi-way assessment of outcomes.

2) *Dataset curation and creation:* Construction of a new, segmented 9000 glyph dataset that will be made publicly available. To our knowledge, this will be the largest public database of individual Maya glyphs.

3) *Glyph representation:* Assessing traditional shape descriptors and representations transferred from deep convolutional networks in a glyph classification task. Different settings in the classification task illustrate the challenges of the new dataset. We also mapped glyphs into 2-D space based on their shape representations.

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 glyph area (overlapping 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, the baseline classification experiments show that the standard transfer learning approach from deep convolutional networks is promising even in the case of few examples per class (around 80% average accuracy in 150-class case). The adopted transfer learning approach with VGG-16 network outperforms traditional shape descriptors by a large margin (around 22% to 37% absolute improvement).

The rest of the paper is organized in eight sections. Section II 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
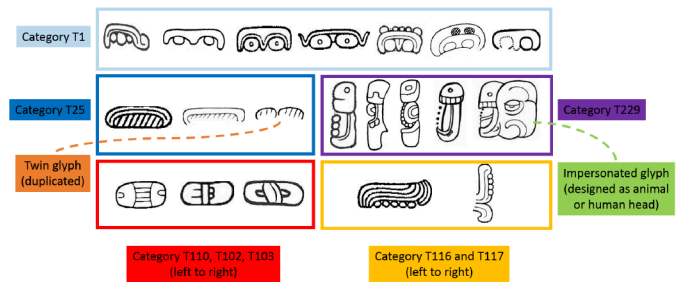

Fig. 2: Selected Maya glyph samples from several categories that illustrate the within-class variety (first two rows) and between-class similarity (last row). Glyph images are provided by Carlos Pallán Gayol.

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 of the crowdsourcing task. Section VIII presents the baseline glyph classification results obtained on the dataset resulting from the crowdsourcing task. Finally, Section IX concludes the paper.

II. MAYA WRITING

The ancient Maya civilization (around 2000 BC to 1600 AD) left a great amount of cultural heritage materials, such as 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. A *glyph* is a unit sign of the Maya writing. A *glyphblock* is composed of several glyphs. A typical *page* of a *codex* is composed of many glyphblocks structured as a grid, and some other pictorial elements. A *codex* is composed of several such pages. 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 having this segmentation task performed by non-experts with carefully-designed support.

The main challenge of our task lies in the nature of the data. Some glyphs are damaged or have many variations due to space limitations, artistic reasons, and the evolving nature of language, i.e., differences with respect to the era and place in which glyphs were produced. Fig. 2 shows the variations of some glyphs in the top two rows.

Another challenge of our study is lack of data, since there are only three genuine codices today. Table I shows the available elements in each of these codices. Among these codices, the shape variation of the glyph categories is relatively low. 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. Since these monument examples are dominant in the glyph catalogs [47], [34], [35], it is difficult to recognize the codex glyphs by just training a model on the catalog examples or monument glyphs. These points motivate us to prepare a crowdsourced glyph segmentation task.

III. Related Work

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

Crowdsourcing in Multimedia and Computer Vision. Several widely-used benchmarks have been produced via crowdsourcing for recognition, detection, segmentation, and attribute annotation tasks. These large-scale datasets enable to train more capable models in multimedia and vision[43], [32].

Crowdworkers motivated by monetary rewards (in crowdsourcing platforms) as well as volunteers have been 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 [21], [22], [28] and astronomy [17].

Task Design. Gottlieb et. al. discuss the key elements in designing crowdtasks for satisfactory outcomes, even for relatively difficult tasks [20]. 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 workers might not have been exposed to such visual data. In order to guarantee satisfactory outcomes, the verification step may require an expert.

Crowdsourcing in Digital Humanities. Digitization and transcription of historical documents with the help of crowdworkers is a widely-studied task in Digital Humanities. 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 [51]. Several decades of the New York Times' archives have been digitized in this way. In similar transcription tasks [10], [9], 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 preliminary work [7], we investigated the perception of glyph shape by non-experts, e.g. whether they saw closed contours as a separate glyph, or how they combined 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
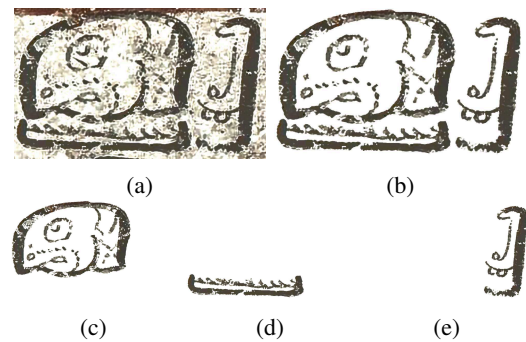

Fig. 3: The top row shows a cropped glyph-block (B1 from fifth page and second t'ol of the Dresden codex) and its cleaned image. The bottom row shows the individual glyphs in the block. These are produced by experts.

variants that guide non-experts to produce arbitrary shape segmentations, and use it to segment over 9000 individual glyphs.

Glyph and Shape Recognition. For Maya glyph recognition, several shape representations have built upon traditional knowledge-driven descriptors [41], [26]. These representations are based on bag-of-words (BoW) that output the frequency histograms of local shape descriptors. As shown in a similar study on Egyptian glyphs [18], HOOSC [41] was a competitive candidate among other traditional shape descriptors.

On the other hand, for shape encoding with neural networks, a single-layer sparse autoencoder, which encodes the same local regions as HOOSC, was shown to be competitive for 10-class monumental glyph classification task [8]. However, this shallow representation was not representative enough for other tasks, i.e. the sketch classification task proposed in [14]. Due to the scarcity of the strokes in thin sketch drawings and the high variety of the drawings, the BoW frequencies of the simple edge encodings in the shallow sparse encoder were harder to capture than thicker glyph strokes. Complementary to this finding, the "Sketch-a-Net" [54] illustrated that a modified version of the AlexNet (in multiple scales and multiple temporal channels) can achieve high performance on the 250-class sketch dataset of [14]. This model has fewer feature maps, yet larger first layer convolution kernels compared to the AlexNet [29], which is designed for natural images.

In the context of Maya glyph-block retrieval, Roman-Rangel showed that the middle-layer activations (conv5) of VGG [46] outperform both the last-layer activations (fc-7), and the bag-of-words representation of a traditional shape descriptor (HOOSC) [40]. This is a motivating point for learning the representations for Maya glyphs, and taking advantage of existing pretrained networks.

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 also use the glyph signs from the Thompson and Macri-Vail catalogs. The details of these datasets are given below.

TABLE I: The number of elements in the three codices (DRE: Dresden, MAD: Madrid, PAR: Paris).

	# 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

A. Maya Codex Glyphs

Our sources are high-resolution digital images from the three existing Codices (Dresden [1], Madrid [2], and Paris [3]), cropped to smaller units (pages, t'ols, and glyph-blocks), and annotated with metadata. Images and annotations were all provided by project partners in epigraphy. The metadata of each glyph-block contains the name of the codex, page number, t'ol number, reading order, and relative location in the t'ol (row and column order, i.e., A1, B2, etc.). The metadata of each glyph in each glyph-block contains its reading order, its sign code from various catalogs (Thompson [47], Macri-Vail [35], Evrenov [15], and Zimmermann [56]), 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.

Table I summarizes the number of elements available from the three Codices. Some pages of the Codices are highly damaged. Even though there are, respectively, 76, 112, and 22 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 epigraphy experts have not provided the individual glyph images for all these glyphs, as the segmentation of Codices into individual glyphs is 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. 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 them [26]. At the large scale, the experts provided only the cropped block images (as in Fig. 3a) without binarization. The details of this raw glyph-block dataset are documented in [25]. Therefore, in order to obtain the individual glyph regions in the blocks, we designed a segmentation-oriented crowdsourcing task.

B. Catalog Signs

The documentation of the ancient Maya writing started during the Spanish conquest of Yucatan in the XVIth century. The first incomplete alphabet [12], [49] was created by asking two locals how to write Spanish characters in Maya language [52]. In the 1960s, Evrenov's [15] and Thompson's [47] 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. With the advancement of the understanding of the semantics of the signs, more modern catalogs emerged [34], [35].

The Thompson catalog has three main categories: affix, main, and portrait signs. Macri-Vail taxonomy has 13 main categories [35]. Six of them (animals, birds, body parts, hands, human faces, and supernatural faces) are grouped semantically. There is a main category for numericals signs that are composed of dots and bars. The rest are grouped based on visual elements (square signs divided based on symmetry, and elongated signs divided based on the number of components).

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 are spread over three Thompson categories (T534 main sign, T140 and T178 affix signs).

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 the partners in epigraphy. As some of the numeric signs were missing in these catalogs, we manually generated them by combination of dots and lines from existing number signs.

Utilizing these variants in a crowdsourcing task has not been previously attempted. Gathering crowd-generated assessments of the similarity between glyph variants and codex glyph samples is 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. We 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 randomly-chosen 10 blocks) with two different task designs. This stage helped to define a final task design. The small stage consists of segmenting glyphs that have ground truth (a subset of glyphs from [26]). This stage helped to judge which catalog was more helpful to non-experts in our task. At the large stage, we conducted the segmentation task for over 10K glyphs.

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 final 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 crowdworkers to segment each individual sign in a block. As Maya glyphs can be found in articulated forms, i.e. hand signs, cropping glyph regions via bounding boxes may end up with inclusion of some parts of the neighbor glyphs. Therefore, for better localization, we designed the segmentation process to be done as free-polygons rather than bounding boxes.

To guide the process, we show workers the different variants of the sign to be segmented. As validation information, we would like to know what sign variant the annotator chose as template to segment each glyph, and how similar the chosen variant and the marked region. This can be used to verify the expert annotations and detect outliers, in case when 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 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. This helps to obtain a damage score that is not affected by prior expert knowledge. The score can also be used as a hint to assess the task difficulty.

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

B. Platform

Terminology. We utilized the Crowdflower (CF) platform for our experiments. In 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 that should 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 dataset to be annotated. The job owner 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 the *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. If 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. The 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 whose overlap with the ground truth region is below the IU threshold, the contributor fails the test question and cannot take on more tasks in the job. 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 answers.

The platform 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 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, 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.

C. Preliminary Stage: Design Experiences.

In the preliminary stage, we conducted four 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. 4) 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 proved to be 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 to the single glyph case. More specifically, the f-measure of segment S and ground truth G is defined based on precision p and recall r as follows:

$$f = 2 * \frac{p * r}{p + r} \quad , \quad p = \frac{|S \cap G|}{|S|} \quad , \quad r = \frac{|S \cap G|}{|G|}. \tag{2}$$

Number of glyph variants. We limited the number of glyph variants shown to the contributors to keep them focused on the

TABLE II: Preliminary stage segmentation results using variants of Thompson catalog (T).

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

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 more visual cues about the possible variations, we decided to provide a maximum of six variants (if available).

Design of feedback mechanisms. In the initial design, we asked contributors about glyph damage level as well as wrong or missing annotations. This part was often omitted by the workers. From this experience, we decided to keep only the most direct rating factors (damage and task difficulty). We also included a text box for optional comments. Received comments included remarks about rotations of the glyph variants, uncertainty about the damage rating, and choice of the variants. 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 the number of tasks per page to 2, set the payment per page to $0.30, and only allowed expert contributors (level-3). This resulted in 79.9% saved segmentations with average f-measure of 79.5%. Considering that there are three glyphs in glyph-blocks in average, we set the payment to $0.10 for the last two single glyph-based experiments to maintain payment/time ratio. Together with the simplified design and the introduction of test questions, this payment and level of expertise 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 it, we decided to collect fewer judgments of 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 workers from different channels (CF-elite channel compared to other channels) in the last two experiments. With the simplified individual glyph-based design, and with level-3 contributors, we did not experience a significant difference in the segmentation scores from these separate channels (89.7%

Fig. 4: Initial block-based task design, illustrating only the first glyph in the block for brevity. Glyph variant images are provided by Carlos Pallán Gayol.

vs. 92%, see Table II). As a consequence, we decided to use all the channels in the following stages.

D. Final Task

1) Overview: Based on the outcome of the preliminary stage, we designed the final task comprising two parts (Fig. 5).

Fig. 5: Final task design. Glyph variant images are provided by Carlos Pallán Gayol.

In the first one, based on the shown variants, contributors were asked to segment (draw a tight free-hand polygon) 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 to rate how similar the variant was to the segmented region, how damaged the glyph region was, and how easy it was to complete the task. These ratings are designed on a scale between 1 and 5.

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

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 increase quality.

VI. EXPERIMENTAL PROTOCOL

Given the decisions made during the preliminary stage, we first conducted 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.

TABLE III: Experimental settings for the small-scale stage (S-1 and S-2) and the large-scale stage (L-1 and L-2).

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

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 with Thompson (T) and Macri-Vail (MV) references of the glyphs. In other words, we display the glyph variants from either the Thompson or the Macri-Vail catalogs.

In both cases, the number of judgments N_j was set to 5. The minimum acceptable IU score was set to 0.7. The minimum time to be spent on a page was set to 30 seconds. The maximum number of judgments by a single contributor was set to 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. We decided to exclude the following glyphs from the annotation:

• Too damaged glyphs according to the damage scores by 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 to be annotated (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 (5000 glyphs) and 0.8 for the rest. This threshold ensured that the contributors did a good job on the test questions, 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 minimum IU threshold for the second half of the glyphs. The minimum time spent on the task
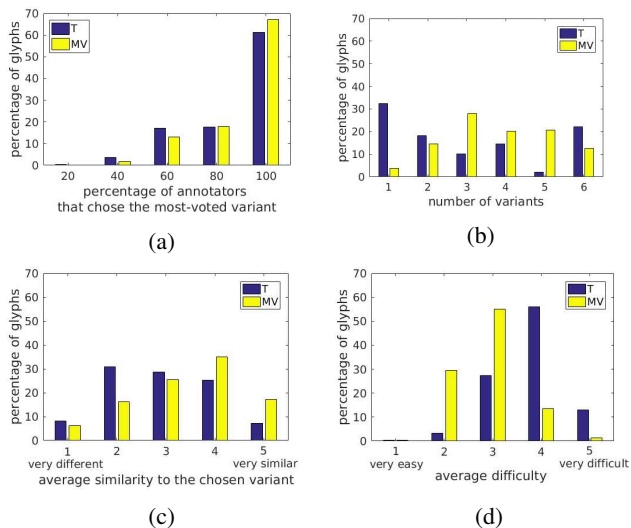
Fig. 6: Distributions of average ratings in the small-scale stage with Thompson (blue) and Macri-Vail variants (yellow).
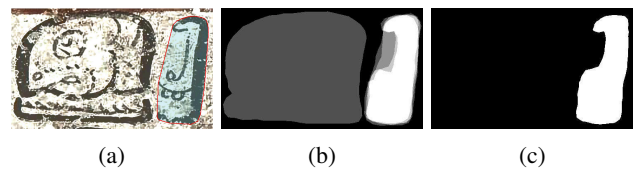

Fig. 7: (a) Convex hull of the ground truth for the glyph on the right (red line, blue filling), (b) gray-scale image of the aggregated segmentations, and (c) final aggregated segmentation.
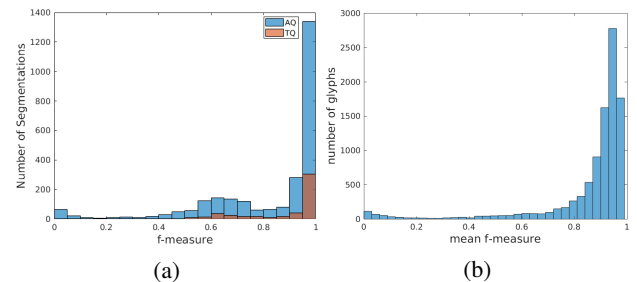

Fig. 8: (a) The f-measure distributions of overlap between crowd segmentations and ground truth in actual question set (AQ, blue) and test question set (TQ, orange) with the MV variants in the small-scale stage. (b) The mean f-measure agreements for the glyphs in large-scale stage.

was set to 30 seconds. The maximum number of judgments by a single contributor was set to 48.

C. Segmentation Evaluation Procedure

Evaluation was 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 problematic cases in which the f-measure agreement was less than 0.8 among contributors as an internal task in Crowdflower platform.

VII. CROWDSOURCED ANNOTATION ANALYSIS

In this section, the crowd annotations for the 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, with Thompson (T), and with Macri-Vail (MV) references of the glyphs. We analyze the annotations from these experiments w.r.t. 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 was selected by majority voting among the contributors' responses. Fig. 6a 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 67.2% of the glyphs when the MV variants (yellow) were shown (61.2% for the T case).

Fig. 6b 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. Thus, even though there were in general more variants available, for the MV cases full agreement was higher (Fig. 6a).

A related result is illustrated in Fig. 6c. Contributors gave higher ratings of visual similarity to the MV variants rather than T variants (2.98 vs. 2.46 mean similarity). Moreover, the contributors found the task harder in the case of T variants (Fig. 6d). These differences in similarity and difficulty ratings were significant as measured with Kolmogorov-Smirnov non-parametric hypothesis testing [36].

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

2) Damage Rating: The average damage ratings (scale 1 to 5) by 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 1 to 5). However, the damage perception of the non-experts was focused around the middle of the scale. For 64% of the glyphs, the contributors selected "moderate-damage" (3 in the range 1 to 5) for both T and MV cases. This can be interpreted as the raw block crops being visually noisy in most of the cases, even though for the experts the glyphs are in good conditions to be identified.
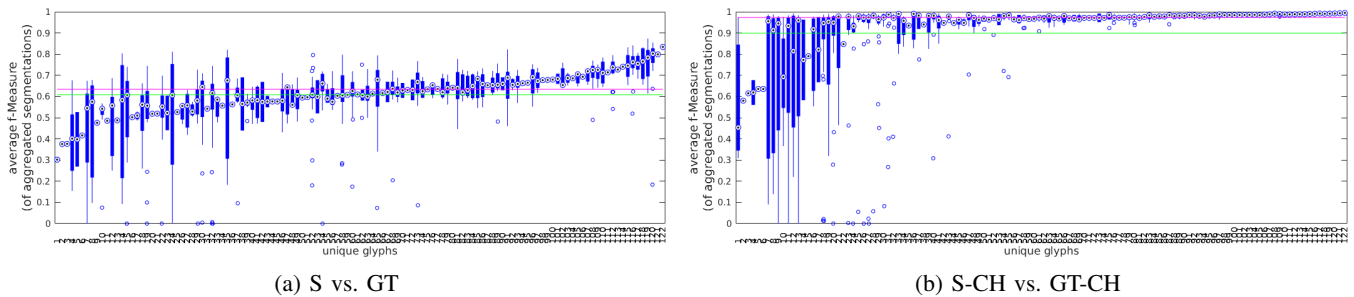
(a) S vs. GT

(b) S-CH vs. GT-CH

Fig. 9: Sorted average f-measure of aggregated segmentations for the unique glyph categories in the small-scale stage. Green and red lines indicate overall mean and median values, respectively.

TABLE IV: Average f-measure values of aggregated segmentations obtained with Thompson (T) and Macri-Vail (MV) variants in small-scale stage for test questions (TQ) and actual questions (AQ).

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

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. 7.

The evaluation is performed by comparing (1) the aggregated segment against the binary ground truth (S vs. GT); and (2) 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 with very detailed contours. Therefore, we decided to use convex hulls for further evaluation in Figs 8-9.

Table IV summarizes the comparative segmentation performance with the help of the two catalogs. It is observed that the MV variants helped to bring out marginally better aggregate 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. 8 (the T variants case is similar and thus not shown). 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 higher mean f-measure for TQ compared to AQ.

Fig. 9 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). 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. When using the convex hull comparison, only ten classes have an average f-score less than 70%.

4) Sensitivity to The Number of Annotators: We simulated the performance for the case of fewer annotators. Fig. 10 shows the average f-measure values for the aggregated masks with different number of segmentations (2-5). 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. The standard deviation of the f-measures obtained with randomly sampled 2-annotations are below 0.1 and are usually acceptable. These observations motivated us to perform the large-scale stage with two annotations per glyph and validate the segmentation when the agreement was higher than 0.8.

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 5 and 6, respectively). This evaluation shows that the defined task is simple enough for a non-expert to produce satisfactory results. Even though the contributors may get confused, overall the performance was high enough to proceed with the large-scale stage.
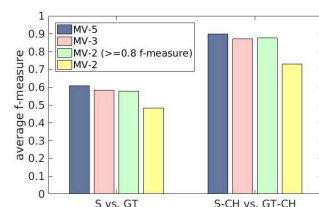

Fig. 10: Mean f-measure values of the aggregated masks obtained using 5 (blue), 3 (pink), 2 (yellow) segmentations, and 2 segmentations that have at least 0.8 f-measure agreement (green) per glyph with MV variants.
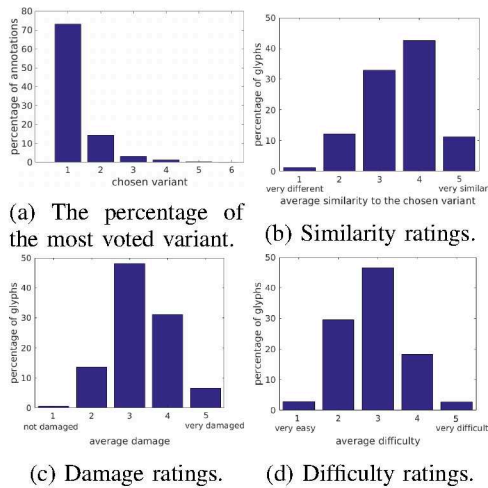
(a) The percentage of the most voted variant.

(b) Similarity ratings.

(c) Damage ratings.

(d) Difficulty ratings.

Fig. 11: Distributions of the ratings in the large-scale stage.

B. Large-Scale Stage

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

1) Glyph Variant Selection: Fig. 11a 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 drawings of more complex monumental glyphs taken from the Macri-Looper catalog [34]. In 7.7% of the annotations, the "none of the variants" option was chosen.

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." The reason behind it may be the tendency of workers to be conservative 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, the general tendency of the contributors (41.9% of the annotations) was to give an average score. However, there are still cases marked as "damaged" or "very damaged" (30.6%), even though we provided glyph cases that are in good condition according to the experts. We believe that workers give 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 positive feedback from the crowd about the perception of the task complexity.

3) Segmentation Analysis: Fig. 8b shows the overall f-measure agreement distribution for the large-scale set.

Verification. In this step, we inspected the segmentations to spot problematic cases. For the cases with f-measure agreement above 0.8, there was a small portion of glyphs (318 out of 8229), in which both contributors marked another region as the glyph area. In the cases with low agreement (1991 glyphs with f-measure below 0.8), we checked if the individual segmentations were usable. In these ways, we exploited all the possible useful segmentations.
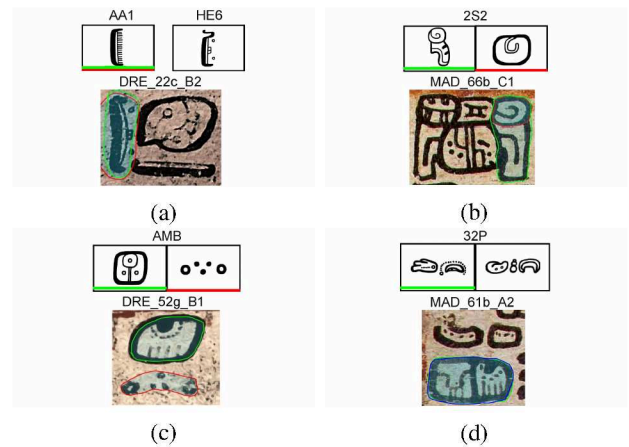

Fig. 12: Confused segmentations from the large-scale stage due to (a) similar glyphs in the block, and damaged instances, (b-c) visually-confusing variants, (d) dissimilar glyphs. Red and green colors indicate the markings of the first and the second worker, respectively.

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 ground truth of the test questions was set to 0.7. This threshold was 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%). Overall, the obtained segmentations are of high quality.

Challenging Cases. The difficulty of our task is not uniform across the glyph instances. Fig. 12 illustrates some of the cases with high disagreement between segmentations. The main reasons for 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. In Fig. 12c, one contributor selected a concave large glyph (green) somehow resembling the first variant instead of the red target region.
- Confusion due to variants: Some variants are a subset or superset of others (i.e., 2S2), as shown in Fig. 12b.
- Dissimilarity between the target region and the variants: We identify three subcases.
  - Target sample not covered by catalog variants. In Fig. 12d, the target region is missed by all contributors, and the neighboring glyphs were marked instead.
  - Partial dissimilarity of the glyph. Some glyphs exhibit partial elements different to the variants (Fig. 12b).
  - Wrong class annotation. In the 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 experts and non-experts due to different use of context or visual completeness. In Fig. 12a, none of the contributors marked the target region, as the target region is either damaged or lacks partial details.
- Similarity to other glyphs in the block. In Fig. 12a, even though the target glyph belongs to class AA1, not HE6, the

TABLE V: The number of glyphs for the classification tasks.

		Number of classes				
		10	**30**	**50**	**100**	**150**
Number of samples	min	211	83	50	20	5
	mean	255.7	176.16	132.66	81.19	57.74
	median	234.5	172.5	101	49.5	26.5
	total	2557	5285	6633	8119	8661

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 33). This stage produced satisfactory outcomes with two non-experts per sample and minimal manual verification. Overall, we obtained valid segments for 9119 glyphs (together with the ones from the small-scale stage) that are spread over 291 MV categories, with the average f-measure agreement 0.914. Most of these valid segments (8661 out of 9119) belong to the most frequent 150 classes in our dataset. We used these aggregated valid segments in the classification task described in the next section.

VIII. BASELINE CLASSIFICATION EXPERIMENTS

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

A. Data Preparation

Our goal is to define a baseline method that highlights challenges and possible classification tasks for our dataset. To assess the difficulty of our dataset, we experimented with different number of classes (the most frequent ones). We considered glyphs with at least one valid segmentation. We have 11 classes with more than 200 such glyphs, whereas 52 classes have just one such glyph. Table V shows the number of glyphs for each experimental setting (the maximum number is 384). For each glyph, to obtain a square crop centered on the aggregated binary mask, we applied the following steps.

- Dilation: We dilated the aggregated mask in case of segmentation not covering all boundary pixels. We set the dilation 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 $0.65 * threshold$ using Otsu's method [38]. In the need of padding, we filled the areas with these RGB values. Note that this step 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 dilated aggregated mask. Then, we padded the short edge to make the final crop square-sized.
- Scaling: We scaled all processed square crops to 224×224 pixels.

After these preprocessing steps, we shuffled and divided each set of glyphs to training (60%), validation (20%), and test sets (20%) for five folds. We report the average accuracies among the 5-folds.
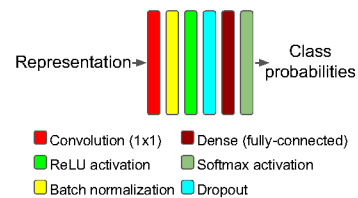

Fig. 13: The shallow CNN model for classification of the representations.

1) Sampling Strategy: We refer to *sampling* as selecting a predetermined number of data samples. In this context, we use the term of *original sampling*, when we employ all the available samples as shown in the last row of Table V. Furthermore, to handle the data imbalance issue among the categories, we considered undersampling and oversampling as alternative strategies. For undersampling, we randomly picked the same number of samples per each class in each experiment (based on the minimum numbers in Table V, 200, 80, 48, 20, and 5, respectively). For oversampling, we applied random geometric data augmentation, comprising rotation (within $[-15, 15]$ degrees), vertical and horizontal translation ($\pm 0.1 \times$ image width), and zooming (scale within $[0.8, 1.2]$). We oversampled the existing examples such that each class had 1000 training, 300 validation, and 300 testing samples. Therefore these oversampled sets were a mix of original data and synthetic data.

B. Methodology

To assess the shape representations for glyph recognition tasks, we evaluated (a) two traditional shape descriptors, i.e. the bag-of-words representation of a local shape descriptor (HOOSC) [41], and a multi-level HOG [11], and (b) the knowledge transfer approach from different pretrained networks [13], [44]. We describe each of these methods below.

1) Traditional Shape Descriptors: For the bag-of-words on the HOOSC descriptors, we followed the same pipeline as proposed in [26] with an additional normalization factor at the end. The steps are as follows.

HOOSC Descriptor Extraction. After binarizing the glyph segments via global Otsu's method [38] (threshold is determined on the corresponding glyph-block image), and applying morphological operations (i.e. closing), we obtain the glyph skeletons. Skeletons are used to select pivot points, and we compute the HOOSC descriptor around each pivot point. To define the local neighborhood while computing the HOOSC descriptor, we used 2-rings and the whole glyph context. The HOOSC descriptor around a pivot point counts the normalized frequencies of the skeleton points in two radial circles (8 orientations), and quantize them in 8 bins. This process produces a 128-dimensional local descriptor around each pivot point. We did not consider concatenating relative spatial location of the pivots here. We randomly selected 400 pivots or more $(0.1 * N_{skeletonpoints})$ from each glyph skeleton if possible, otherwise we used all the skeleton points as pivots.

After extracting the local HOOSC descriptors for each glyph, we sampled 80% of the glyphs randomly. From this

TABLE VI: Average classification accuracies on the original sets with a linear SVM (S) and the shallow CNN (N) in Fig.13.

	Original Sampling									
	Number of classes									
	10		30		50		100		150	
Model	S	N	S	N	S	N	S	N	S	N
HOOSC	70.1	69.8	57.4	57.8	49.5	50.1	44.0	43.1	39.7	40.3
HOG	67.2	71.1	52.8	56.8	46.0	50.3	41.8	44.5	39.2	41.4
SaN_B	81.6	85.7	70.5	76.7	63.5	71.6	58.2	66.0	56.1	63.4
SaN_RGB	84.4	88.6	74.7	81.0	70.2	77.0	65.2	73.0	62.5	70.1
VGG16	**92.0**	**91.8**	**89.9**	**89.0**	**86.6**	**84.2**	**82.6**	**82.3**	**80.0**	**79.2**
R50	75.7	81.7	65.4	72.9	51.8	68.1	46.0	63.2	41.5	59.5

Fig. 14: Partial visualization of the 2S2 glyphs via t-SNE algoritm shows the separation of glyphs corresponding to two different variants (see Fig.12b, blue cluster for the first, pink cluster for the second variant).

set of glyphs, we sampled 10% of the HOOSC descriptors of each glyph to build the dictionary by applying k-means with 4000 cluster centers.

After computing the dictionary with vocabulary size 4000, we assign each HOOSC descriptor of each glyph to their closest cluster center (or word in the dictionary) with $L1$ distance. Therefore, for each glyph, we obtain a codebook that corresponds to the frequencies of closest words of its HOOSC descriptors in the dictionary. The final representation HOOSC-BoW has 4000 dimensions.

Multi-Level HOG Descriptor Extraction. We concatenated the histogram of orientation features at two-levels. We computed the HOG with 13×13 and 24×24 pixels cell sizes and 4 blocks in each cell with 9 orientations. Since our images have 224 pixel image size, we ended up with $16 \times 16 + 8 \times 8 = 320$ cells, and $320 * 4 * 9 = 11520$ feature dimension for each image.

Normalization. Due to the nature of the BoW computation, i.e. hard-assignment, the HOOSC-BoW representation is distributed among the 4000 dimensions with a constraint on the dimensions summing up to 1. A normalization of this representation with a scaling factor is needed to obtain a reasonable comparison with CNN activations. Therefore, we first normalized the BoW vectors of each glyph with the corresponding max value, i.e. making the max value of each vector equal to 1, and then scaled the BoW vectors with a constant scalar to match the maximum activation value of the pretrained CNN features. A similar normalization is applied to the HOG features.

Classification. The HOOSC-BoW and multi-level HOG features are used as input to a shallow neural network (Fig. 13) with two fully-connected (FC) layers. The first FC layer has 1024 filters. We applied ReLU activation between two FC layers as well as batch normalization [27], and dropout [24] method with 0.5 rate. The final class probabilities are determined by the softmax activation at the end. Additionally, we assessed the representations with a standard linear support vector machine (SVM) as well.

2) Pretrained CNN Features: CNNs pretrained on large-scale datasets, i.e. ImageNet, are used as feature extractors by feedforwarding the image of interest, and gathering the activations at different layers of the network [13], [44], [53], [48], [39], [55]. The penultimate activations before softmax classifier have been reported as good baselines for transferring knowledge in several vision tasks [13], [44]. Furthermore, the middle-layer activations are more generic than the last-layer
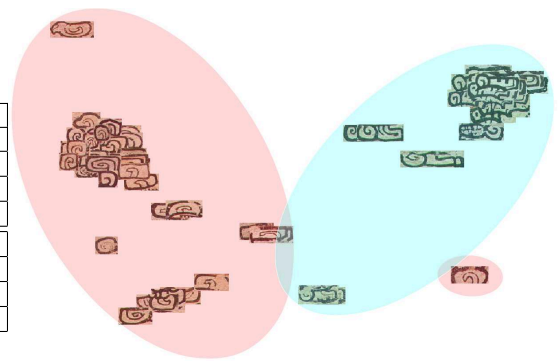
ones, and may be more applicable to the data with different nature (e.g. man-made vs. natural objects) [53].

With this motivation, we forward the glyph segments in our dataset through a pretrained network, and collect the activations at the end of the last convolutional block. We consider these activations as our pretrained CNN features.
Considered Networks. We considered the VGG-16 network [46] and ResNet-50 [23] pretrained on ImageNet dataset, and the Sketch-a-Net [54] pretrained on 250-class binary sketch images [14].

VGG-16 is a 16-layer CNN model, shown to be competitive on the ImageNet dataset before the inception module and residual connections were introduced. We passed our RGB glyph images from the pretrained VGG-16, and extracted the activations from the last (5th) convolutional layer. Similarly, for the ResNet50, we extracted the activations from the last global average pooling layer (just before the FC layer and softmax classifier).

Sketch-a-Net (SaN) is adapted from the AlexNet model [29] for handling sparse sketch images. We retrained the single-scale single-channel version of the SaN model: adding batch normalization (BN) layer [27] after each convolutional and dense layer. The modified SaN obtained competitive results on a random split of the sketch dataset (72.2% accuracy). We used this model to extract the activations of the binarized version of our glyph images. Similarly, we retrained another SaN with the fake-colored sketch images (filled with same RGB values that are used to populate our glyph dataset). We passed our glyph

TABLE VII: Average accuracies on the original test sets for pretrained features, when the shallow CNN networks were trained on the undersampled vs. oversampled sets.

| | | Number of classes | | | | |
	Model	10	30	50	100	150
Undersampling (on training)	SaN_RGB	87.9	76.6	67.6	54.6	29.1
	VGG16	91.3	84.8	78.0	64.1	35.2
	R50	79.4	63.8	51.4	35.9	16.5
Oversampling (on training)	SaN_RGB	95.6	93.0	91.5	90.0	71.4
	VGG16	97.0	96.1	95.0	93.6	80.6
	R50	93.5	90.2	88.2	86.1	62.0

images (either binary or RGB) through these networks, and extracted the activations from the $6th$ convolutional block. To assess these representations, the same classifiers were applied as noted in Section VIII-B1.

C. Classification Results

Table VI shows the average accuracies among 5-fold experiments with original sampling in different settings. 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 91.8% average accuracy with the VGG-16 pretrained features. For the 150-class case, we obtained 79.2% accuracy (random guess would be 0.66%). Table VI confirms the competitiveness of the pretrained CNN features, that are learned from large-scale datasets, compared to traditional shape descriptors. Among the pretrained net features, the VGG-16 activations provide the best results. Furthermore, Table VII points out that oversampling during training helps all the models and improve over undersampling with a large margin. These results both show the challenges and complexity of our dataset and encourage further work in the future.

D. Visual Analysis with t-SNE

To further understand the characteristics of the curated dataset, we mapped the segmented glyph samples to a 2-D space in terms of visual similarity (obtained via methods mentioned above). This mapping is realized via the t-distributed Stochastic Neighborhood Embedding (t-SNE) [50]. The visualization enables to see all the samples of the same category that are scattered in a quantitative manner. This visualization could help to assess the glyphs in the "gray areas" (highly-discussed with scholars in terms of identification), as the glyphs are mapped to a visual similarity context. This visualization can also help experts in catalog design, as the main variations of the sign categories are clustered together thanks to this mapping.

Fig. 14 presents a visualization of the set of segmented glyphs from the 2S2 class displayed via t-SNE algorithm over the last convolutional layer activations from the Sketch-a-Net pretrained on 250-class-sketch data. In this example, it is interesting to notice the separation of the glyph instances corresponding to the different variants.

IX. FINAL CONCLUSIONS

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

- **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 a simpler and focused task design (to segment individual glyphs rather than all glyphs in a block) and clear instructions were indispensable.

- **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 non-experts to reach a higher consensus on the "closest-looking" variant, and obtain higher agreement (average f-measure). Furthermore, we observed that workers found the task easier with MV variants. These results were to some degree expected as monumental glyphs were the main source of Thompson catalog variants.

- **Non-expert behavior analysis.** We pointed out the main challenges that workers faced during the task, such as visual within-class dissimilarities or between-class similarities, and the effect of damage. These challenges affect the segmentation outcome. However, they are inherent to the data.

- **Maya codical glyph corpus.** This work generated over 9K individual glyphs from the three 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 that transfer learning methods with deep neural networks are promising.

REFERENCES

[1] "Dresden Codex," http://digital.slub-dresden.de/werkansicht/dlf/2967/1/.
[2] "Madrid codex," http://www.famsi.org/mayawriting/codices/madrid.html.
[3] "Paris Codex," http://gallica.bnf.fr/ark:/12148/btv1b8446947j.
[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] ——, "Evaluating shape representations for Maya glyph classification," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 3, sep 2016.
[9] L. Carletti, G. Giannachi, D. Price, and D. McAuley, "Digital humanities and crowdsourcing: An exploration," in *Museum and the Web*, 2013, pp. 223–236.
[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] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.
[12] C. E. B. de Bourbourg, *Relation des choses de Yucatan de Diego de Landa*. Durand, 1864.
[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition."
[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] M. Franken and J. C. van Gemert, "Automatic Egyptian hieroglyph recognition by retrieving images as texts," in *International Conference on Multimedia*. ACM, 2013, pp. 765–768.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[25] R. Hu, G. Can, J.-M. Odobez, and D. Gatica-Perez, "The Maya codex glyph block dataset," Idiap, Tech. Rep. Idiap-Internal-RR-34-2017, May 2017.

[26] 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.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of International Conference on Machine Learning*, 2015, pp. 448–456.

[28] 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.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in NIPS*, 2012, pp. 1097–1105.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

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

[39] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[40] E. Roman-Rangel, G. Can, S. Marchand-Maillet, R. Hu, C. Pallan Gayol, G. Krempel, J. Spotak, J.-M. Odobez, and D. Gatica-Perez, "Transferring neural representations for low-dimensional indexing of Maya hieroglyphic art," in *ECCV Workshop on Computer Vision for Art Analysis*, Oct. 2016.

[41] E. Roman-Rangel, C. Pallan, J.-M. Odobez, and D. Gatica-Perez, "Analyzing ancient maya glyph collections with contextual shape descriptors," *IJCV*, vol. 94, no. 1, pp. 101–117, 2011.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

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

[48] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

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

[50] L. van der Maaten and G. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *JMLR*, vol. 9, pp. 2579–2605, 2008.

[51] 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.

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

[53] 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.

[54] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-net that beats humans," *arXiv preprint arXiv:1501.07873*, 2015.

[55] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[56] 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. He is a member of the IEEE, and Associate Editor of the IEEE Transaction on Circuits and Systems for Video Technology and Machine Vision and Application journals. Email: odobez@idiap.ch

Daniel Gatica-Perez (S'01, M'02) is the Head of the Social Computing Group at Idiap Research Institute and Professor Titulaire at EPFL, Switzerland. He has served as Associate Editor of the IEEE Transactions on Multimedia. He is a member of the IEEE. Email: gatica@idiap.ch