Is That a Jaguar? Segmenting Ancient Maya Glyphs via Crowdsourcing

Gulcan Can
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne, Switzerland
gulcan.can@idiap.ch

Jean-Marc Odobez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne, Switzerland
odobez@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne, Switzerland
gatica@idiap.ch

ABSTRACT

Crowdsourcing is popular in multimedia research to obtain image annotation and segmentation data at scale. In the context of analysis of cultural heritage materials, we propose a novel crowdsourced task, namely the segmentation of ancient Maya hieroglyph-blocks by non-experts. This is a task that is highly perceptual and thus potentially feasible even though the crowd is not likely to have prior specialized knowledge about hieroglyphics. Based on a new data set of glyph-block line drawings for which ground-truth segmentation exists, we study how non-experts perceive glyph blocks (e.g. whether they see closed contours as a separate glyph, or how they combine visual components under plausible hypotheses of the number of glyphs present in a block.) Using Amazon Mechanical Turk as platform, we perform block-based and worker-based objective analyses to assess the difficulty of glyph blocks and the performance of workers. The results suggest that a crowdsourced approach is promising for glyph-blocks of moderate degrees of complexity.

Keywords

crowdsourcing; Amazon Mechanical Turk; Maya glyphs

1. INTRODUCTION

Image labeling via crowdsourcing has been extensively used to generate large amounts of labeled training data, necessary for object detection, recognition, and segmentation tasks in computer vision and multimedia [12, 2, 10, 11]. Strategies based both on games [12] and monetary rewards using platforms like Amazon Mechanical Turk have demonstrated their utility in producing labeled image sets of adequate reliability for a variety of generic content labels in natural images, including objects, actions, and scenes [6].

From a different angle, crowdsourcing has been successfully used to produce linguistic resources of historical and cultural heritage materials, using e.g. the re-captcha paradigm to transcribe old documents, using a combination of automated document analysis methods and human intelligence [13]. Several decades of the New York Times archives have been digitized in this way. Crowdsourcing is also finding other applications in the digital humanities [1].

We are interested in developing crowdsourcing techniques to support the development of multimedia analysis methods for ancient Maya hieroglyphics present in digital images from a variety of vestiges including monuments, stones, ceramics, and codices. Our work involves a tight collaboration with scholars in Maya and archeology [8, 9].

The ancient Maya writing system is complex and highly visual. Typical inscriptions contain iconography (drawings representing deities and people) and hieroglyphs structured in so-called glyph-blocks. Glyph-blocks are typically composed of a small set of individual glyphs. Single glyphs can correspond to syllables (syllabograms) or concepts (logograms).

One basic pressing need is the generation of segmented and labeled glyph data to train and test machine learning methods [3]. Given that the ancient Maya language has no similarities to other existing languages today, segmented glyphs are typically produced by experts. This is a very time-consuming task, and often tedious for highly trained scholars. On the other hand, glyphs are at their core *visual patterns* and often resemble known objects like animals, human body parts, etc. One could wonder whether the general human ability to recognize visual patterns could be used for a relatively simple task, namely locate individual glyphs within a glyph-block, with no previous training. In other words, given a single glyph-block and using only perceptual information, could people guess the number of glyphs and draw bounding boxes around them? If feasible, this could provide a cost-effective alternative for collecting annotation labels for simple tasks.

In this paper, we investigate whether reliable annotations of non-experts can be generated as a crowdsourced, glyph-block segmentation task. To our knowledge, this question is novel both in computer science and in digital humanities. For this, we developed an interactive interface and used Mechanical Turk as platform. We use a new data set of glyph-block line drawings for which ground-truth segmentation exists in terms of number of glyphs and their location, which allows to objectively assess the performance of non-experts. We use best practices in Mechanical Turk (regarding requirements for workers and monetary incentives) to recruit workers, controlling for an inherent measure of task complexity (the number of glyphs in the block N_b). Based on the crowdsourced results involving both a pilot study with known workers and a full study with mTurk workers, we

CrowdMM'14, November 07 2014, Orlando, FL, USA.

ACM 978-1-4503-3128-9/14/11...$15.00.
http://dx.doi.org/10.1145/2660114.2660117.

show that the task is feasible for glyph-blocks of moderate visual complexity (defined by N_b), and that visual complexity has a clear effect on segmentation performance, measured objectively w.r.t. to the ground-truth. Our framework is overall promising.

2. OVERVIEW OF OUR APPROACH

We conduct two studies, a pilot study and a Mechanical Turk (mTurk) study. In both studies, during the annotation task, for a given block, workers have to provide (1) the segmentation of each glyph as a bounding box, (2) a perceived number of glyphs in the block, and (3) the rating of the task difficulty. The annotations are analyzed with respect to: 1) task difficulty, 2) range of perceived number of glyphs and 3) segmentation performance by comparing the number of bounding boxes and their location with the ground truth. Accuracy and purity measures of the crowdsourced segmentations are examined both block-wise and worker-wise.

3. DATA DESCRIPTION

The ancient Maya civilization flourished from BC 2000 to 1500 AC in Mesoamerica. Mayan art can be found in stone monuments, codices pages, and ceramics. As explained in Section 1, the Maya writing system is composed of glyphs. They are generally structured in blocks where several glyphs come together and are meant to be read in a specific order.

In this work, we used line drawings generated from stone monuments in Yaxchilan, an archaeological site located in the state of Chiapas in Mexico. The data consists of drawings of glyph blocks present in monuments, depicting the visual content with high fidelity. In order to keep the annotation task feasible, we have selected glyph blocks having 3, 4 or 5 glyphs. Note that this range accounts for the majority of blocks in the data sets we currently work with and constitute a measure of visual complexity.

Segmentation of glyphs in these blocks can be quite challenging for non-experts due to erosion, occlusions, and the inherent visual richness of the glyphs themselves. In this work, we have not used severely eroded blocks. Figure 5 illustrates three block examples. The leftmost column corresponds to the ground truth, and from top to bottom, 3-, 4- and 5-glyph examples can be observed. We use a total of 50 glyph-blocks, 31, 12, 7 for 3-, 4-, and 5-glyph cases.

4. CROWDSOURCING TASK

We developed a user interface for bounding box annotation, comprising three parts: training, drawing, and evaluation.

Training. To train the workers, we provide clear guidelines, a how-to video, and examples for each category (please see http://youtu.be/WDEmubaF2x0). The how-to video gives a brief introduction to the Maya writing system, and how to use the interface. To be clear about the task, we also provide a few positive and negative examples. Obviously, bounding boxes covering very small areas are not desired. Negative examples also include cases of too-much-overlap and not-enough-image-coverage. Our goal is that after these guidelines, workers will rely on their perceptual skills.

Drawing. In the main drawing pane, the worker clicks on one edge to start drawing a bounding box and ends by clicking on the diagonal edge. The worker can also remove bounding boxes. The main pane also provides information about the expected block complexity expressed as a range for the number of glyphs in the blocks. This is a key piece of prior knowledge to focus the human task on a narrower set of possible answers. At the same time, it reflects the natural statistics of glyph-blocks.

Evaluation. We also ask the workers to rate the difficulty of segmenting the block (in a scale of 5) and to declare the approximate number of glyphs they would have provided if we had not specified it a priori, namely less than 3, between 3 and 5, and more than 5.

Worker Population. Given the novelty of the segmentation task, we decided to first conduct a pilot study with a smaller set of glyphs and workers we personally knew before launching the mTurk study. The first pilot study has 15 participants and 30 glyph-blocks, whereas in mTurk study there are 10 annotators per block and 50 glyph-blocks. In the pilot task, 3-, 4- and 5-glyph blocks have the same number of examples (10) each. However, in the mTurk study, as we have selected blocks with catalog annotations, the number of glyphs per block category is 31, 12 and 7 respectively. In the pilot task, participants are not paid, however they are committed and reliable sources. In the mTurk study, we limited our crowd to the ones with *master*'s level expertise and an acceptance rate of at least 95%. In terms of time required to collect the annotations, for the pilot study it took approximately 10 hours to get responses from all participants, whereas for mTurk study it took around 2 hours. The estimated task duration is around 1 minute. Each mTurk HIT was paid at 0.15 USD.

5. RESULTS AND DISCUSSION

In this part, we analyze the crowdsourced data from three perspectives: task difficulty, glyph range perception, and segmentation performance. For the pilot study, we have 450 annotations for 30 blocks; for the mTurk study, 500 annotations for 50 blocks from 23 unique workers.

Task Difficulty. For this analysis, the explicit ratings of the workers about the drawing task are evaluated. Figure 1 plots the relative proportion for 3-, 4-, and 5-glyph cases. Interestingly, in the mTurk study, workers tend to mark the task easier than in the pilot study. On the other hand, as the number of glyphs increase, the increasing trend of hard and very hard ratings remains similar in both studies. We can conclude that 3-glyph cases are considered easier than 4-glyph and 5-glyph cases.

Range Analysis. For this question (number of glyphs

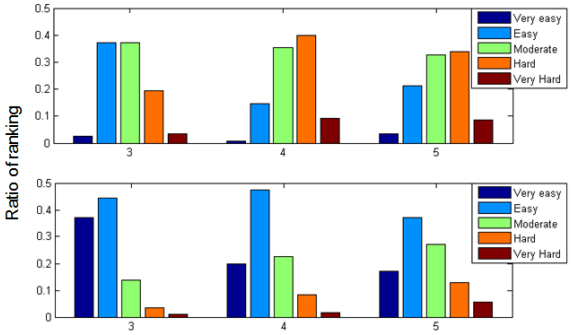

Figure 1: Task difficulty from pilot study (top) and mTurk study (bottom) for 3-glyph, 4-glyph, and 5-glyph blocks.

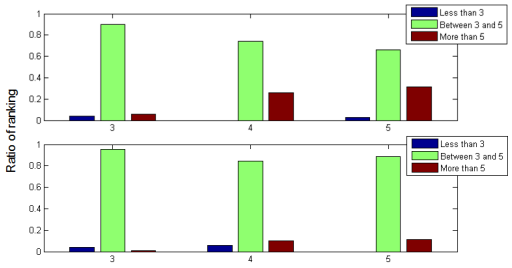

Figure 2: Proportion of perceived number of glyphs from pilot study (top) and mTurk study (bottom) for 3-glyph, 4-glyph, and 5-glyph cases.

guessed without any constraints), the distributions are plotted in Figure 2. Although it can reasonably argued that people were biased towards the suggested range, workers still choose out-of-range options and the plots still indicate a decreasing trend of "in the range of 3-5" as blocks get more complex, more noticeably in the pilot study.

Segmentation Performance. Segmentation annotations are studied in two aspects: number of bounding boxes and area-wise comparison of segmented vs. ground truth bounding boxes.

Bounding Box Number Analysis. As observed from Figure 3, there is a decreasing trend in the correct number of bounding boxes as glyph complexity increases. This is expected, since people get more confused about marking more complex glyphs. Interestingly, the mTurk workers did a better job for the 3-glyph case (0.8 vs 0.6).

Area-Based Performance Analysis. To measure the objective performance of the bounding box annotations, two metrics (accuracy and purity) are used:

$$accuracy(A,G) = \frac{1}{N_o}\sum_k max\left(\frac{|a_k \cap g_{j_k}|}{|a_k \cup g_{j_k}|}\right) \quad (1)$$

$$purity(A,G) = \frac{\sum_k max_j |a_k \cap g_j|}{\sum_k |a_k|} \quad (2)$$

where $A = \{a_1, a_2, ..., a_k, ..., a_n\}$ is the set of the annotation bounding boxes of a worker for a glyph-block, and $G = \{g_1, g_2, ..., g_k, ..., a_n\}$ is the set of ground truth bounding boxes for that glyph-block. Correspondence between an annotated bounding box a_k and a ground truth bounding box g_j is found by $j_k = \underset{j}{argmax}\left(\frac{|a_k \cap g_j|}{|a_k \cup g_j|}\right)$. N_o stands for the number of annotated boxes who suffice an overlapping

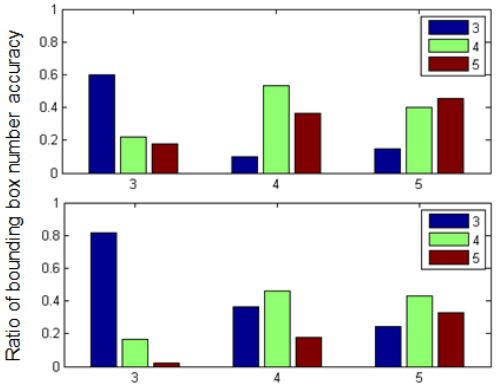

Figure 3: Percentage of bounding boxes from pilot study (top) and mTurk study (bottom).

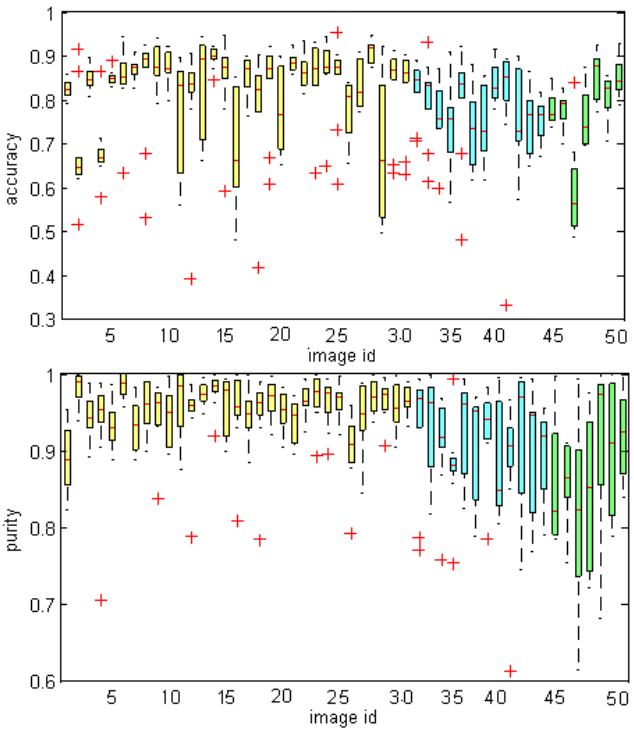

Figure 4: Block-based annotation accuracy (top) and purity (bottom) from mTurk study. Yellow: 3-, blue: 4-, green: 5-glyph blocks.

constraint.

For accuracy, the mean intersection over union ratio of annotation and ground truth bounding boxes is computed. With this measure, we penalize sloppy annotations. Equation 2 is the well-known cluster purity measure [7] defined over bounding box regions. These two measures are correlated by a factor of 0.61 in the mTurk data.

Block-based Analysis. In the mTurk study, high performance values are obtained, however the mean values decrease and the standard deviation increases as blocks get more complex (see Table 1). Figure 4 shows the accuracy and purity of mTurk annotations. As Table 1 and Figure 4 show, blocks with fewer glyphs are segmented more accurately, with highest values of 0.82 and 0.95 for accuracy and purity for the 3-glyph case.

In Figure 5, the first row shows the best case of anno-

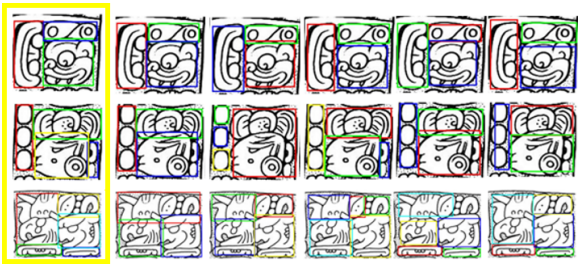

Figure 5: Top two and bottom scored image annotations from mTurk. The first column is the ground truth, and other columns are the annotations of 5 workers. Drawings produced by Graham and Von Euw © [5, 4], block segmentation and glyph annotations provided by Carlos Pallán Gayol. Visualize in pdf for details.

Table 1: Block-based annotation performance for mTurk study for 3-glyph, 4-glyph, and 5-glyph cases.

	Mean Acc.	Std. Acc.	Mean Pur.	Std. Pur.
3	0.820	0.063	0.951	0.020
4	0.725	0.047	0.907	0.022
5	0.692	0.103	0.871	0.042

tation based on accuracy values (where annotators guessed correctly). The other rows show the worst annotations. The bottom row is a 5-glyph block where workers get confused about the glyphs on the top as well as whether to merge the small elongated glyphs on the lower part with the head-like shapes on top of them. We can also see this merging tendency issue of elongated glyphs with the head-like glyphs in the second example. We can also observe that worker 2 marked three circles on the left separately, probably because they are well separated, and mark the rest complex part as one.

About the annotations, we observe that they are in good quality in general. For instance in the bottom row, only worker 4 has not left an unmarked closed contour on the upper right part. We encountered very few sloppy bounding boxes and no random marking at all. We hypothesize that the coverage and overlap constraints on the user interface helps increase the high performance values.

Worker-based Analysis. Performance for each worker is shown in Figure 6, where workers are ordered based on the number of blocks they annotated (shown as percentages on top of the bars). We observed that some workers marked only a few blocks, which is typical in crowdsourcing. Their performance is sometimes better than the few workers who worked almost all of the blocks as the latter must have encountered hard cases in the dataset as well (and possibly experienced fatigue). Average accuracy per worker ranges between 0.64 and 0.92 as purity is between 0.88 and 0.98.

6. CONCLUSIONS

We presented a new use of crowdsourcing for generating segmentations of ancient Maya glyphs by non-experts. The task was designed as a constrained segmentation problem with little prior training, and that largely relied on perceptual organization skills of workers. Using a variety of segmentation quality measures, we conclude that the task is feasible for moderate visual complexity (measured by the number of glyphs in a block), and that less complex blocks (containing 3 glyphs) were indeed easier than other cases.

Given the formidable challenges of the Maya script, by no means we claim that the crowd can substitute expert knowledge in epigraphy. Rather, the results suggest that non-expert work could be useful for simple, well-designed segmentation tasks, which could later be verified by experts. In the future, in addition to using a significantly larger data set, we will investigate whether more accurate segmentations can be obtained with variations of the task presented here, e.g. by modifying the interaction paradigm or adding information coming from extra sources like glyph catalogs.

Acknowledgments. This work was funded by the SNSF MAAYA project. We thank our partner Carlos Pallán Gayol (University of Bonn) for providing the glyph data, Darshan Santani and Laurent Nguyen (Idiap) for help with the crowdsourcing task; Rui Hu (Idiap) for discussions; and all workers for their participation.

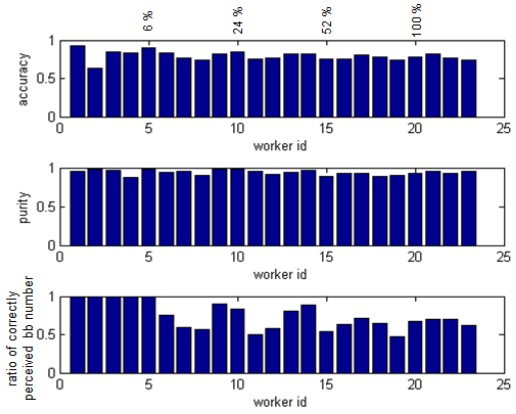

Figure 6: Worker-based accuracy (top), purity (middle), and ratio of correctly perceived glyph number (bottom) from mTurk study.

7. REFERENCES

[1] L. Carletti, G. Giannachi, and D. McAuley. Digital humanities and crowdsourcing: An exploration. *Museums and the Web, Portland, Oregon.*, 2013.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, 2009.

[3] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proc. ACM SIGIR*, pages 871–880. ACM, 2012.

[4] I. Graham. *Corpus of Maya hieroglyphic inscriptions*, volume 3. Peabody Museum of Archaeology and Ethnology, Harvard University, 1979.

[5] I. Graham and E. Von Euw. *Corpus of Maya hieroglyphic inscriptions*, volume 3. Peabody Museum of Archaeology and Ethnology, Harvard University, 1977.

[6] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. Jones. The community and the crowd: Multimedia benchmark dataset development. *MultiMedia, IEEE*, 19(3):15–23, July 2012.

[7] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.

[8] E. Roman-Rangel, C. Pallan, J.-M. Odobez, and D. Gatica-Perez. Analyzing ancient maya glyph collections with contextual shape descriptors. *IJCV*, 94(1):101–117, 2011.

[9] E. Roman-Rangel, C. Pallan Gayol, J.-M. Odobez, and D. Gatica-Perez. Searching the past: an improved shape descriptor to retrieve maya hieroglyphs. In *Proc. ACM Multimedia*, pages 163–172. ACM, 2011.

[10] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[11] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820, 2008.

[12] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. ACM CHI*, 2004.

[13] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.