Evaluating Shape Descriptors for Detection of Maya Hieroglyphs

Edgar Roman-Rangel[1,⋆], Jean-Marc Odobez[2,3], and Daniel Gatica-Perez[2,3]

[1] University of Geneva, Switzerland
[2] Idiap Research Institute, Switzerland
[3] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
edgar.romanrangel@unige.ch,
{odobez,gatica}@idiap.ch

Abstract. In this work we address the problem of detecting instances of complex shapes in binary images. We investigated the effects of combining DoG and Harris-Laplace interest points with SIFT and HOOSC descriptors. Also, we propose the use of a retrieval-based detection framework suitable to deal with images that are sparsely annotated, and where the objects of interest are very small in proportion to the total size of the image. Our initial results suggest that corner structures are suitable points to compute local descriptors for binary images, although there is the need for better methods to estimate their appropriate characteristic scale when used on binary images.

Keywords: Shape detection, image retrieval, Maya hieroglyphs.

1 Introduction

The interpretation of ancient Maya inscriptions requires the identification of the basic individual components (glyphs) of the Maya writing system. Currently, this identification process is performed manually by experts, who often need to consult printed catalogs [1], [2]. However, often the size of the individual glyphs is considerably small in proportion to the size of a complete inscription, thus making laborious the manual detection process.

The complexity of the manual detection process increases if we take into consideration that, it is a common feature of the Maya writing system to arrange glyphs at arbitrary position within the inscriptions. Therefore, the implementation of techniques for automatic detection of these complex glyphs requires special attention. Fig. 1 shows and example of a Maya inscription.

One issue related to the automatic detection of Maya hieroglyphs, is that currently, the amount of annotated data that is available remains limited, thus making difficult the implementation of supervised learning methods.

In this paper we present the results of an evaluation of shape descriptors for the automatic detection of Maya glyphs, using weakly annotated binary

⋆ This work was conducted at Idiap as part of the first author's doctoral research.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2013, LNCS 7914, pp. 145–154, 2013.

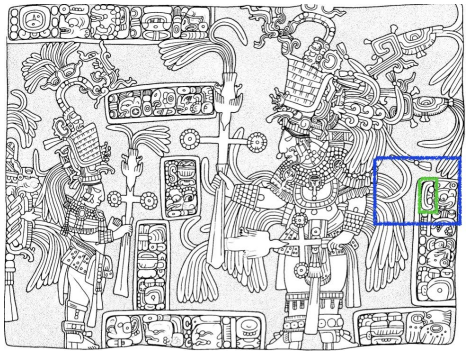


Fig. 1. Maya inscription from the archaeological site of Yaxchilan, Mexico. In a green rectangle, there is an instance of syllable u used as ground-truth for the random block bounded by the blue rectangle. © Image provided by archaeologist Carlos Pallan (University of Bonn), we use this image with his consent.

images. We believe this work will have a positive impact on the daily work of archaeologist diminishing the time required to decipher Maya inscriptions. Namely, the contributions of this paper are:

1. The generation of a synthetic dataset of Maya syllables, which was gathered to overcome the issue of only having few available instances in each visual class. Because of the nature and visual complexity of its instances, this is a unique and highly valuable dataset.
2. The approximation of the sliding-window detection approach by a retrieval-based detection scenario. Our approach overcomes the issue of having only few annotated data, which constrains the use of supervised learning methods.
3. The evaluation of two popular interest point detectors, and their combination with two state-of-the-art image descriptors on the task of shape detection. More precisely, the DoG [3] and Harris-Laplace [4] interest point detectors, and the SIFT [3] and HOOSC [5] descriptors.

Note that the HOOSC descriptor builds on top of Shape Context (SC) [10], and that was proposed to overcome some of its limitations when dealing with shapes that are more complex than the brand logos SC was evaluated on.

The rest of this paper is organized as follows. Section 2 discusses the related work in shape description and image detection. Section 3 introduces the dataset we used in this work. Section 4 explains our experimental protocol. Section 5 discusses the results obtained with our approach. Finally, in section 6 we present our conclusions and a discussion on the open issues.

2 Related Work

The representation of shapes is a research topic with long tradition [6], [7], [8], [9]. In a nutshell, shape descriptors differ according to whether they are applied

to contours or regions, and whether they describe global or local patterns of the shapes. For instance, descriptors based on moments are relatively easy to compute, and they are robust against location, scale, and rotation variations [6]. And Fourier descriptors work well for simple shapes of convex contours [7]. However, both of them perform poorly with affine transformations, and for complex shapes whose instances have many local variations. Also, they need efficient approaches to normalize descriptors derived from different shape signatures [8].

Shape context descriptors [10] incorporate robustness against affine variations, and are able to deal with shapes of high visual complexity [5]. However, the size of the bounding box containing the shape of interest is of high relevance for the normalization, which in principle is unknown on a detection setup. Therefore, they are not suitable for detection purposes.

Several approaches have shown success in the task of detecting objects on gray-scale images [11]. The common framework for image detection implements a sliding-window, in which a classifier is used to evaluate sub-windows and decides whether or not they contain the element of interest. However, such methods require having enough amount of data to train the classifier. Another limitation for using traditional gray-scale oriented approaches [3], is that they rely on local regions of interest whose size is estimated using the information provided by local intensity changes [4], and this information is absent in binary images.

Common approaches to deal with the problem of detecting shapes address these issues by relying on shape information estimated upon gray-scale images, i.e., by extracting contours and local orientations based on the local gradients of intensity images rather than using binary images [12]. For instance, using a networks of local segment as descriptors, and performing detection of shapes belonging to classes that are relatively easy to differentiate in visual terms [13].

In contrast, in this work we address the problem of detection of complex shapes that exist as binary images. These shapes belong to visual classes that exhibit high levels of both inter-class similarity and intra-class variability, thus making the problem more challenging. Also, we implement an ad-hoc approach to address the issue of having limited amount of data to train a classifier.

3 Dataset

We use blocks randomly segmented from very large inscriptions to have a better control over the experimental setup. The reason for this is that the inscriptions are very sparsely annotated relatively to their size and content, such that there is a high probability of detecting non-annotated true-positive instances. With this purpose, we generated three set of images: ground-truth, positive, and negative instances.

More specifically, we followed a five-steps process for data generation: (1) First, we chose 24 visual classes of syllabic Maya hieroglyphs, and for each of them, we manually located 10 different instances on a large collection of inscriptions (thus, 240 instances). We labeled this set as *ground-truth* instances. The reason to choose only 24 visual classes is because they correspond to the hieroglyphs that were most commonly used, thus facilitating their manual location

and segmentation. (2) Then, we generated a random block for each ground-truth instance. This generation of random blocks consisted in segmenting a sub-window containing the ground-truth itself plus a surrounding area, with the restriction that the left and right margins surrounding the ground-truth had a random size between one and four times the width of the respective ground-truth, while the top and bottom margins have random sizes between one and four times the height of it. The decision to use such values is to generate random blocks that contained enough visual information around the ground-truth, such that the challenge of a realistic detection setup is kept. Fig. 2 shows the details of the random block highlighted in Fig. 1. (3) The next step consisted in annotating the random blocks, such that the bounding box of each ground-truth is known, relative to the random block and not to the original large inscription, i.e., once a random block was segmented, we annotated the coordinates x and y where the ground-truth bounding box starts, and its corresponding width (w) and height (h). (4) Later, we generated 20 variants of each ground-truth by randomly shifting the position of its bounding box up to 0.2 times its width and height respectively, and we annotated the location (x, y, w, h) of these variants. This resulted in 200 instances per syllabic class that we labeled as *positive*. (5) Finally, for each segmented random block, we annotated the location (x, y, w, h) of all the existing bounding boxes that are of the same size as its respective ground-truth, but that do not overlap with it. This last part resulted in 6000+ bounding boxes that we labeled as *negative* instances. On average, each block contributed with 26.1 ± 16.0 negative instances.

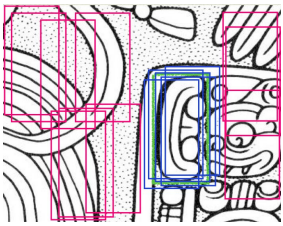


Fig. 2. Random block extracted from the inscription shown in Fig. 1. The ground-truth corresponds to an instance of syllable *u*, and it is inside the green rectangle, the positive instances are marked with blue rectangles, and the red rectangles indicate some of the negative instances.

In summary, the dataset is composed by 24 syllabic classes, in total containing 240 ground-truth instances (10 for each syllabic class), 4800 positive instances (200 for each syllabic class), and 6000+ negative instances that do not belong to any of the positive classes. By annotating the images in this way, we turned the traditional detection approach based on sliding-windows into a retrieval-based approach. This change avoids the risk of detecting non-annotated true-positive instances, and resulted in fast detection experiments, although at the price of non-exhaustive scanning of the large inscriptions.

4 Experimental Protocol

This section explains the experimental protocol followed to evaluate the detection performance achieved by using DoG and Harris-Laplace interest point, and combining them with the SIFT and HOOSC descriptors. Table 1 summarizes the combinations we evaluated.

Table 1. Tested combinations of interest points and local descriptors for detection of Maya syllables

Name	Interest points	Descriptor	Input format
DoG-SIFT	DoG	SIFT	shapes with thick contours
DoG-SIFT-thin	DoG	SIFT	shapes with thinned contours
DoG-HOOSC	DoG	HOOSC	shapes with thinned contours
HarrLapl-HOOSC	Harris-Laplace	HOOSC	shapes with thinned contours

For the DoG and SIFT implementations we used the OpenCV libraries, and we implemented the Harris-Laplace and HOOSC methods in Matlab. Since the HOOSC descriptor was developed to deal with medial axes of shapes, and with the purpose of comparing the two descriptors, we also computed DoG points and SIFT descriptors for the thinned versions of the shapes, as shown in Table 1. Namely, we performed our experiments under the following six-steps protocol:

1. **Interest point detection:** First, we detected points of interest (DoG or Harris-Laplace), along with their characteristic scales and local orientations on the random blocks. For the computations of interest points we considered each random block as a whole (i.e., the points of interest were not computed individually per each bounding box), thus avoiding potential boundary effects as in a common detection setup.
2. **Description:** Second, we computed the local descriptors (SIFT or HOOSC) using the point's characteristic scales and local orientations. This computations were also performed over each complete random block.
3. **Estimating visual vocabularies:** After computing the sets of descriptors for all the random blocks, we randomly drew 1000 descriptors (SIFT or HOOSC) from each visual class and clustered them into 1000 "words". To do so we used the k-means clustering algorithm.
4. **Indexing:** Then, we constructed bag-of-visual-words (*bov*) representations individually for each bounding box. The *bov* were constructed taking into account only those points whose characteristic scale was relevant within the current bounding box, thus excluding points that might contain more information about the exterior than about the interior of the bounding box. Therefore, we excluded: (1) points whose scale is much larger than the bounding box, and (2) points near the the edge of the bounding box and whose scale only intersects a small proportion of it. More specifically, we excluded

all those points whose ratio of intersection $r = A/(2s)^2$ was below 0.5, where s is the characteristic scale of the point, and where A is the intersection area between the characteristic scale and the current bounding box.

5. **Detection:** After computing the *bov* representation of each bounding box, we computed the euclidean distance from each ground-truth's *bov* against the *bov*s of all the positive and negative bounding boxes extracted from the random blocks of the same class as the current ground-truth, i.e., we performed detection on weakly annotated random blocks, looking for instances for which we know they are present inside a given random block. Note that we excluded the random block that contains the current ground-truth, as itself and all its positive variants are easily detected. In practice, each ground-truth is expected to have smaller distances to 189 bounding boxes (the other 9 ground-truth instances plus their 180 positive instances) than to the negative instances (on average, 234.8 negative bounding boxes per class). Thus our detection method is not a classical exhaustive sliding-window but an approximation based on a retrieval approach.
6. **Evaluation:** Finally, we ranked all the bounding boxes based on the computed distances, and evaluated the detection performance in terms of,
   - ROC curves. Comparing the mean average detection-rate (mA-DR) versus the mean average false-positive-rate-per-window (mA-FPPW) at various threshold values.
   - Curves showing the average-precision achieved at different top-N positions of the ranked subwindows.
   - The mean Average Precision (mAP).

Note that the HOOSC descriptor, as described in [5], has five main characteristics. Namely, (1) it uses thinned versions of the shapes, (2) estimates local descriptors only for certain locations (termed pivots) with respect to whole set of points in the thinned shapes, (3) computes an histogram of local orientations in each of the regions of a polar grid around each pivot, (4) in turn, the spatial scope of the polar grid is defined as a function of the average pair-wise distance between all the point in the thinned shape, and (5) the explicit relative position of the pivot to be described may be used as a part of its own descriptor.

These characteristics of the HOOSC descriptor work well in tasks such as classification and retrieval of shapes that have been previously segmented and where the instances are not rotated or reflected. However, such assumptions are not true in the case of a detection setup. For our experiments, it was not possible to compute the spatial scope of the polar grid as a function of the pair-wise distances of the contour points, as the correct size of the bounding box is unknown a priori, and evaluating all possible sizes would result impractical. Therefore, we made use of the characteristic scale of the interest point (DoG or Harris-Laplace) at which the descriptor is computed. More precisely, the polar grid we implemented has two local rings with boundaries at 0.5 and 1.0 times the characteristic scale of the interest point. Also note that we did not use the explicit relative position of the pivots in their description, as the size of the

candidate bounding boxes is assumed to be unknown, and also because some elements might be rotated within the inscriptions.

5 Results

The ROC curves in Fig. 3 show that the use of DoG points with thinned contours gives detection rates close to chance, both with SIFT and HOOSC descriptors (green and red curves, respectively). This observation is not especially surprising as binary images lack of intensity information which is the main clue to localize DoG interest points and to estimate their characteristic scale. The motivation to use DoG points in thinned shapes was based on the high frequency of blob structures present in the Maya syllables. However, some times of the DoG interest points correspond to large blob structures that encompass visual information beyond the locality of the glyph of interest, which in practice, were excluded as explained in section 4. This in turn, resulted in poor shape representations.

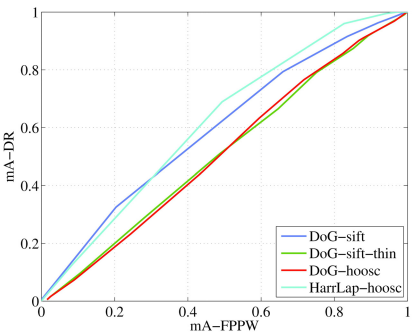
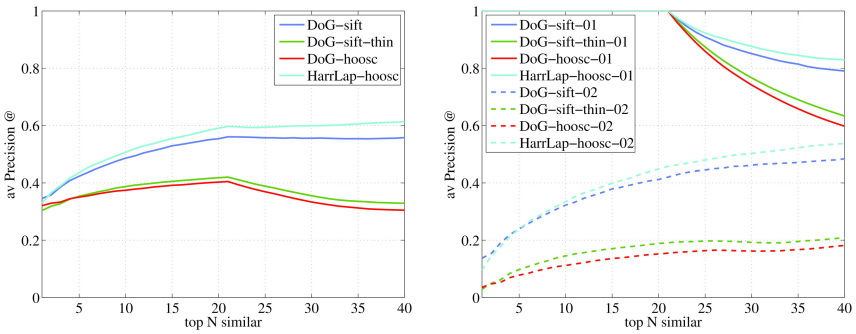


Fig. 3. ROC curves showing the detection performance of different combinations of interest point detectors and shape descriptors

Note that the detection rate is relatively increased with the estimation of DoG points on the original shapes that have thick contours (blue curve in Fig. 3), this is mainly explained by the used of the Gaussian convolutions that smooth the thick contours and approximate intensity values on the resulting image. Moreover, the use of Harris-Laplace interest points resulted in a slightly increased detection rate when used on thinned shapes (see cyan curve Fig. 3).

In terms of retrieval precision, the relative difference among the four methods remains proportional to their ROC curves, as shown in Fig. 4a. The slight peak in the retrieval precision at position 21 results because some classes have instances very similar to one another, such that for a given query (ground-truth instance), the 21 bounding boxes (ground-truth + positive instances) of (at least) one

relevant random block are well ranked at the top of the retrieved vector. To better illustrate this, we recomputed the average precision regrouping the ranked vectors into two sub-groups: one with the queries whose precision curves remain equal to 1 at the 21-st position, and the other with the remaining queries. These results are shown in Fig. 4b, where the solid curves (named XXX-01) show the average precision for the first set, and the dashed curves (named XXX-02) show the average precision of the second set. This said, some visual classes are very easy to retrieve, whereas some others are quite hard.



(a) Average precision at top N **(b)** Split average precision at top N

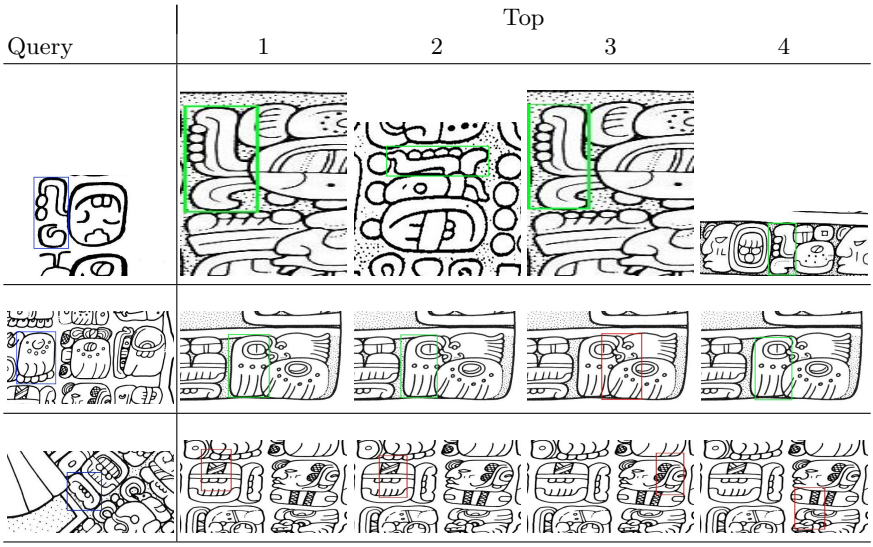
Fig. 4. Average retrieval precision achieved by the different combinations of interest point detectors and shape descriptors. (a) ROC curves, and (b) mean Average Precision curves at top N, plotted for different combinations of interest points and images descriptors evaluated in detection experiments.

The cyan solid curve in Fig. 4b corresponds to HOOSC descriptors computed at Harris-Laplace interest points. Note that this curve remains with good precision values at the 40-th position of the top N vector. Thus indicating that this combination of interest points and descriptor works well in general terms. To summarize the retrieval performance of the tested combinations, we present their mean Average Precision (mAP) in Table 2. Note that the use of corners as interest points achieves better performance than blob structures. Finally, Table 3 shows visual examples of the detection obtained using Harris-Laplace points with HOOSC descriptors.

Table 2. Mean average precision (mAP) for the combinations of interest points and local descriptors tested for detection of Maya syllables

Method	DoG-SIFT	DoG-SIFT-thin	DoG-HOOSC	HarrLapl-HOOSC
mAP	0.614	0.449	0.440	**0.646**

Table 3. Visual examples of detection with Harris-Laplace interest points and HOOSC descriptors. The first random block in each row contains a query inside a blue rectangle (ground-truth). The next four random block correspond to the four most similar bounding boxes according to the method, where green rectangles indicate correct detection, and red rectangles indicate erroneous detection.



6 Conclusions

In this work we explored an initial approach for detection of complex binary images (syllabic Maya hieroglyphs), evaluating the performance of DoG and Harris-Laplace interest points combined with SIFT and HOOSC descriptors.

We presented a controlled retrieval-based framework for detection that can be used as an alternative resource when the data is sparsely annotated, thus avoiding the risk of detecting non-annotated true-positive instances. This setup also avoids the exhaustive scanning of the traditional sliding-window approach.

Our results show that regardless of the local image descriptor, the use of DoG points with thinned contours gives detection rates close to chance as a consequence of the lack of intensity information in binary images. A slightly better performance is achieved by using thicker contours since the Gaussian smoothing approximates some sort of intensity information. Moreover, the use of corner detectors seems suitable for local description of complex binary images, as shown by the detection rates obtained by the Harris-Laplace interest points. In terms of retrieval performance, the HOOSC descriptor achieves competitive results, specially when it is combined with Harris-Laplace interest points.

It is important to remark, that this initial stage suggests the need for interest point detectors specially tailored for binary images, such that regions of interest are located within the shape along with their characteristic scales, and therefore, shape descriptors that have proven successful with segmented shapes can be used also for detection.

Acknowledgments. This research was supported by the Swiss NSF CODICES project (grant 200021-116702). We thank the Swiss NSF through the NCCR IM2 for providing travel funds, and Prof. Stéphane Marchand-Maillet for comments.

References

1. Thompson, J.E.S.: A Catalog of Maya Hieroglyphs. University of Oklahoma Press, Norman (1962)
2. Macri, M., Looper, M.: The New Catalog of Maya Hieroglyphs. The Classic Period Inscriptions, vol. 1. University of Oklahoma Press, Norman (2003)
3. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
4. Mikolajczyk, K., Schmid, C.: Scale and Affine Interest Point Detectors. International Journal of Computer Vision 60(1), 63–86 (2004)
5. Roman-Rangel, E., Pallan, C., Odobez, J.-M., Gatica-Perez, D.: Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptors. International Journal in Computer Vision, Special Issue in Cultural Heritage and Art Preservation 94(1), 101–117 (2011)
6. Hu, M.-K.: Visual Pattern Recognition by Moment Invariants. IEEE Transactions on Information Theory 8(2), 179–187 (1962)
7. Zahn, C.T., Roskies, R.Z.: Fourier Descriptors for Plane Close Curves. IEEE Transactions on Computers 21(3), 269–281 (1972)
8. Zhang, D., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition 37(1), 1–19 (2004)
9. Yang, M., Kpalma, K., Ronsin, J.: A Survey of Shape Feature Extraction Techniques. In: Pattern Recognition, pp. 43–90 (2008)
10. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4), 509–522 (2002)
11. Viola, P.A., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2001)
12. Payet, N., Todorovic, S.: From Contours to 3D Object Detection and Pose Estimation. In: IEEE International Conference on Computer Vision (November 2011)
13. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of Adjacent Contours for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(1), 36–51 (2008)