Chavin de Huantar Potsherd Impression Analysis

Michael P. Kim and Bryce Cronkite-Ratcliff

March 2014

Abstract

Chavín de Huántar is an archaeological site of critical interest for intepreting the Andean archaeological record. In this paper, we present our work on developing an object recognition system to classify Chavín pot sherds by decorative impressions. We trained and evaluated MSER and SURF based classification systems on a small labeled set of replica sherds and achieved classification accuracy of 75%. The performance of our system is comparable with a similar system built for ancient coin classification [8] [9] and represents a promising first step in developing an automatic sherd classification system for archaeologists interested in Chavín pottery or, more generally, in pattern recognition for the analysis of archaeological ceramics.

1 Introduction

1.1 Chavín Pottery

The archaeological site of Chavín de Huántar in the Peruvian Andes has been occupied since at least 3000 BCE and serves a critical role in interpreting the Peruvian archaeological record. In particular, the site appears to have been a primary center of Andean culture for several centuries, commanding loyalty far beyond its borders, and set forth a decorative style that is regularly echoed in subsequent Andean works.

At Chavín, over 18,000 pottery sherds from an estimated 700 distinct pots were unearthed in just one year of excavations. These artifacts were produced by the Chavín culture in the the first millenium B.C.E. and characteristically are decorated by repeated impressions of a stamp into moist clay. The large number of sherds found at the site, as well as other sites throughout the Andes, poses an untapped resource for furthering our understanding of the Chavín. That is, the regularity of the process used to produce each pot should allow association of a pot with a set of particular stamps impressed upon it. The aim of the work here presented is to develop a system to automatcially associate a Chavín sherd with an original set of stamps, thus allowing archaeologists to reconstruct a network of original stamps and excavated pot locations. Such a network may provide insight into the movements of people and

Figure 1: Examples of Chavín Pot Sherds

goods throughout the world of the Chavín, thus allowing for important new insights into the world of the Chavín culture and its mechanisms of cultural dissemination.

1.2 Computer Vision and Archaeological Pottery Analysis

Beyond our interest in Chavín archaeology, there is a broader need for research into computer-aided analysis tools for archaeological artifacts. As the most frequent class of artifacts found throughout the world, pottery (and potsherds in particular) should be of primary focus. It is our belief that recent developments in automated recognition, and in computer vision in particular, have resulted in techniques sufficiently effective in practice to be of great utility to archaeologists. This is in addition to research thus-far in computer vision applications in archaeology, which have focused largely on methods for 3D digital archiving of sites and artifacts and fragment reconstruction. Thus, one research aim of this work is to demonstrate the application of object recognition results to pottery analysis. We are particularly interested in demonstrating that the questions that can potentially be facilitated by such a system extend beyond those of recognition, organization, and archiving, to the recovery of latent information residing in patterns in large bodies of artifacts.

2 Contributions

2.1 Prior Work

Computer Vision applications in archaeology have been applied primariliy to problems of digital archiving. In particular, we see the existing work as falling into three categories:

1. The aquisition of 3D models of sites and artifacts
2. The automatic re-assembly of artifacts from fragments
3. Assisted artifact recall in existing databases of images or models

The first of these includes the work of Geert Verhoeven and colleagues applying GIS to the problem of obtaining 3D site models, applying computer vision techniques in multiple-view geometry, such as Structure from Motion [18] [17] [2]. Work has also been done on the artifact scale on techniques for 3D model building (using techniques such as space carving, range scanning, and photogrammetry) and the related problem of archiving artifacts in a database for recall and analysis [13] [16]. Another angle on 3D modeling of archaeological material is aimed at developing excavation aids – tools to visualize excavation state and systemtize the organization of digital records in a useful manner; this includes work on the Reveal system at Brown as well as work at ETH Zurich [3] [15].

Significant progress has been made in the assembly of artifacts from fragments. This research has taken two approachs, human-assisted and automatic. HINDSIGHT is a recent project from Brown focusing on a complete interactive system for fragment re-assembly [4]. A project out of Stanford attempted reconstruction of the Forma Urbis Romae, a fragmented map of ancient rome, using crowdsourced solving [5]. Research into automatic artifact assembly has been entertained consistently for the past decade [12] [19] [20]. Successful automatic artifact re-assemblies thus-far include a selection of pot fragments from Petra and parts of the Forma Urbis Romae [3]. While this may seem discouraging, Andrew Willis points out that fragment digitization is a major bottleneck in automatic re-assembly [3].

The third area is less explored, but includes work on shillouete similarity metrics for the search of similar artifcats in digital archives using shape context and shock patches [10] [14]. Our work shares the most with a

project seeking to classify ancient coints automatcially to aid ameteur coin classification. This work presents a complicated custom pipeline involving Sobel edge detectors, Gabor wavelets, Shape context detectors, and Principal Compnent Analysis [8]. As a first step to the analysis of ancient coins, the authors classified a dataset of modern coins, reaching 71.5-78% accuracy [8] [9].

2.2 Our Contributions

Our approach differs in that we are interested in applying object recognition tools not primarily for archiving, recall, or database exploration generally, but to answer a specific archaeological question. We not only anticipate that an effective automatic classification system will aid in catalouging and analytics for Chavín sherds, but also believe our work may serve as a model for applying object vision more directly to archaeological problems. Additionally, to our knowledge, the detection techniques we make use of here, such as MSER and SURF, are not used in the archaeological object recognition literature.

Figure 2: Set of Replica Stamps used as our custom dataset

This work is highly experimental. The task of automatically classifying any image of any sherd is too large and involves too many confounding factors to tackle in this report alone. For example, there is the difficulty of automatically identifying individual sherds within an image of multiple sherds and individual stamps on each sherd; the difficulty of variable lighting conditions, clay colors, and light responses (some pots are shiny glazed blackware; some are diffuse bare clay); the difficulty that ground truth is not known; and the issues of large, complex stamps and near-total occlusions where only small fragments of an original impression remain.

In light of this, we chose to focus on a reduced problem for the work presented here. We create an annotated dataset of sherds of one whole impression each and photograph each individually under consistent lighting. Thus, we avoid problems of occlusion, sherd color and light response, multiple registration, and lack of ground truth. We carry out the rest of the impression recognition pipeline on this custom dataset.

3 Technical Approach

3.1 Summary

Our approach can be briefly sumarized as follows:

- Development of custom dataset
- Feature Detection and Extraction
- Classification
- Testing

3.2 Details of Project Components

3.2.1 Development of Custom Dataset

As described in section 2.2, in order to focus on the problem of impression identification without many additional complications, we chose to create our own annotated dataset. We obtined a smooth stoneware clay from the Stanford Ceramic Studio waste clay bin and molded twelve unique stamps. After several days of drying, the stamps were impressed into 2-3 inch diameter round pieces of wet clay to create "sherds." Nine impressions were made from each stamp, and the tweleve stamps were divided into four general "shapes." Each stamp was numbered and labeled, and each impression was labeled with its corresponding stamp by carving the stamp number into the reverse side of each sherd. Thus, the annotated dataset contained in total 108 sherds divided evenly into 12 original stamps or "classes" and divided evenly into 4 general shapes: bad donut, good donut, cross, and crescent. To specify, the dataset contains 9 impressions from each of 12 stamps and thus 27 impressions of each of 4 shapes.

Figure 3: Poloroid Land Camera Photographic Setup with replica stamps

Figure 4: Examples of each Shape: from left to right "bad donut", "good donut", "cross", and "crescent"

Each sherd was placed in a consistent location on the platform of a poloroid Land Camera and photographed at 12 Megapixels using a Panasonic Lumix DMC-GF3K Camera with a 14mm lens(see Figure 3). One of the bulbs on the Land Camera was not functional, so sherds with polar impressions were oriented consistently to obtain consistet shadows. We believe that the missing bulb may have been an advantage, because feature detection algorithms depend on contrast, which was increased by an uneven – though consistent – lighting setup. Each of the sets of 9 impressions associated with the same stamp was photographed together, such

that we can simpy look at the timestamp (or, identicially, the filename) of a image to determine its associated stamp.

3.2.2 Feature Detection and Extraction

Each image was reduced from 12 Megapixels to 3 Megapixels and converted to grayscale to enable processing using color-invariant descriptors. We also created low resolution images of .12 Megapixels. We then separated the data into train and test sets; the first 7 images of each class (in the order photographed) were designated train, and the remaining 2 were designated test. In the following, all tuning and visualization were performed on train images only. We used three established descriptors to detect and extract features: SIFT, SURF, and MSER.

We first applied a SIFT implementation in C by David Lowe and colleagues [6] [7]. We obtained keypoints and used Matlab R2013a to plot the keypoints on the original images. We observed that keypoints clustered around shadow lines and thus occured both within the impression and at the edges of each sherd, as well as on a few scratches on the land camera base.

Using an external system, however, imposed unnecessary infrastructural overhead, so we put aside this SIFT implementation in favor of descriptors built into the Computer Vision Toolbox for Matlab R2013a, in particular SURF and MSER [1] [11].

To start, we tried using SURF features due to their efficiency, robustness, and conceptual similarity to SIFT features. We found that SURF features were fast to compute and generally similar to the features extracted by SIFT. While these features seemed to give robust matches between images, qualitative evaluation of why SURF worked well in some instances but not others was difficult, so we were curious to explore other features as well.

We found MSER intriguing because the implementation we used permitted visualization of the detected regions, which allowed simple qualitative analysis of the utility of a particular feature for describing an impression (See Figure 10). We used this feedback mechanism to tune our detector. We went through several iterations of parameters, but settled on the following for our high resoluton detection:

```
MaxAreaVariation 40, ThresholdDelta 4, RegionAreaRange [400, 8000]
```

3.2.3 Classification

We took two different approaches to classification, Characteristic Feature Response and K-Nearest Neighbors.

Characteristic Feature Response One of the main challenges we faced using any of the featurization methods is how to deal with spurious feature matches. Both feature descriptors tended to discover features in the images that were not part of impressions yet were contributing significantly to matched feature sets. Moving forward we could re-take the images using a chroma key backdrop to reduce suprious feature matches. In order to mimimize the effect of suprious matches with our existing dataset, we performed the following training procedure:

On a training set, we computed the feature matches between all images within each class and compiled them into a list of "characteristic features." In some sense, these features are a good baseline set of features shared amongst images belonging to the class, so might be indicative that a new image belongs in the class. Next, having computed the characteristic features of each class, we computed the matches between different classes' characteristic features. For each pair of classes, we computed matches across the sets of their characteristic features. When features from across classes were matched, we removed these features from

the list of characteristic features, as this cross-class matching indicates that the feature did not distinguish between classes.

We then predict the class of a test image by extracting features and then performing matches against each class's characteristic features. We tested using both MSER and SURF feature descriptors at two levels of resolution. We report the results in the next section.

K-Nearest Neighbors For each image i in the test set, we detected and extracted features in the image and counted the number of matches, M_{ij} found betweeen i and each train image, j. We thus determined a score S_i^K for the similarity of i to each class K as follows:

$$S_i^K = \sum_{j \epsilon K} M_{ij}$$

We first selected the class C_i to which i as follows:

$$C_i = \underset{K}{\text{argmax}}(S_i^K)$$

However, this approach was subject to corruption from discrepancies between the average number of matches seen by images in a given class. For example, train images from class 1 see, on average over matches with all images, about 10 times as many matches as do train images in class 12.

Thus, we decided to normalize based on the distribution of number of matches with images in each class. We compare all train images to each other as described above. Once we have recorded each M_{ij}, we find the sample mean, $\hat{\mu}_K$, and sample standard deviation, $\hat{\sigma}_K$ of all M_{ij} for a particular $K = \{j | j \epsilon \; class \; K\}$. The score of a test image i for class K, denoted $\hat{S}_i^K$ now becomes the Z-score of its M_{ij} statistic:

$$\hat{S}_i^K = \frac{M_{ij} - \hat{\mu}_K}{\hat{\sigma}_K}$$

and the class C_i thus simply becomes:

$$\hat{C}_i = \underset{K}{\text{argmax}}(\hat{S}_i^K)$$

We found this approach to significantly improve results, from near baseline random at 9% classification accuracy for classes to 33%. Results for this method are presented in section 4.

3.2.4 Testing

We tested our results on three full pipelines: MSER with Characteristic Features, SURF with Characteristic Features, and MSER with K-Nearest Neighbors. Our test set consisted of 24 images, two from each class. Recall that ground truth was known because the dataset was built for this work. While our initial pipeline took about 90 minutes to run, by caching extracted features we are able to bring test image running time down to the order of minutes. We ran our tests on Matlab R2013a with the Computer Vision Toolbox on a 2011 iMac with 2.5Ghz I5 and 4GB RAM.

4 Experimental Results and Analysis

4.1 Classification by K-Nearest Neighbors

We tested on 24 images, 2 from each class (labeled i and ii in figure 5) as described above. The optimal results obtained are presented in figure 5. Note that the baselne classification accuracies are about 8% and 25% by class and shape respectively.

True Class	1	2	3	4	5	6	7	8	9	10	11	12
Predicted Class on (i)	1	12	3	11	4	10	7	4	11	12	11	12
Predicted Class on (ii)	1	10	5	12	5	10	10	5	1	6	12	12

Figure 5: 33.0 % accuracy by class / 75.0% accuracy by shape

Note that, distinct from the Characteristic Features approach, this classification method returns both class and shape classifications. We are clearly much more effective at classifying by shape than class, although we beat random meaningfully on within-shape class classification assuming correct shape classification.

4.2 Classification by Characteristic Features

Initially, we learned characteristic features by impression class, but predictions were noisy as few features were selected for each class after cross-class filtering. Thus, we instead learned characteristic feature by shape. The models trained according to the shape appeared more robust. Below, we include the confusion matrices from 4 different runs of the characteristic features classifier on the test set using SURF and MSER features on low resolution (.12 Megapixels) and higher resolution (3 Megapixels) images. In each matrix A, A_{ij} is the number of images of the ith shape that were classified as shape j.

$$\begin{vmatrix} 4 & 0 & 2 & 0 \\ 0 & 4 & 1 & 1 \\ 0 & 0 & 4 & 2 \\ 0 & 0 & 2 & 4 \end{vmatrix}$$

Figure 6: Low Resolution Prediction by shape using SURF (62.5% accuracy)

$$\begin{vmatrix} 4 & 1 & 1 & 0 \\ 0 & 6 & 0 & 0 \\ 5 & 0 & 1 & 0 \\ 0 & 2 & 1 & 3 \end{vmatrix}$$

Figure 7: Low Resolution Prediction by shape using MSER (58.3% accuracy)

$$\begin{vmatrix} 5 & 0 & 0 & 1 \\ 0 & 4 & 0 & 2 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 2 & 4 \end{vmatrix}$$

Figure 8: High Resolution Prediction by shape using SURF (70.8% accuracy)

From the perspective of accuracy, we can see that these classifiers do significantly better than random, and that performance seems to improve with increased resolution. In particular, when we consider the

$$\begin{vmatrix} 6 & 0 & 0 & 0 \\ 1 & 3 & 0 & 2 \\ 0 & 1 & 5 & 0 \\ 0 & 1 & 1 & 4 \end{vmatrix}$$

Figure 9: High Resolution Prediction by shape using MSER (75.0% accuracy)

performance of the classifiers which use MSER features, we see that the classifier using low resolution images consistently labels images of the "good donut" shape as the "bad donut" shape. We see though when we increase the resolution of the images used for training and evaluation, the precision increases and the classifier differentiates between the two types of donut impressions.

It is important to note that these results indicate the performance of the classifiers after tuning the feature extraction and matching parameters. The runs using MSER features were particularly sensitive to approriate specification of the parameters affecting the size of the features detected; we suspect that on the low resolution images, features were lost because they shrunk below our feature detector's minimum region size threshold. In future work, it may be worthwhile to tune detection parameters individually for different datasets, even simply differently scaled versions of the same dataset.

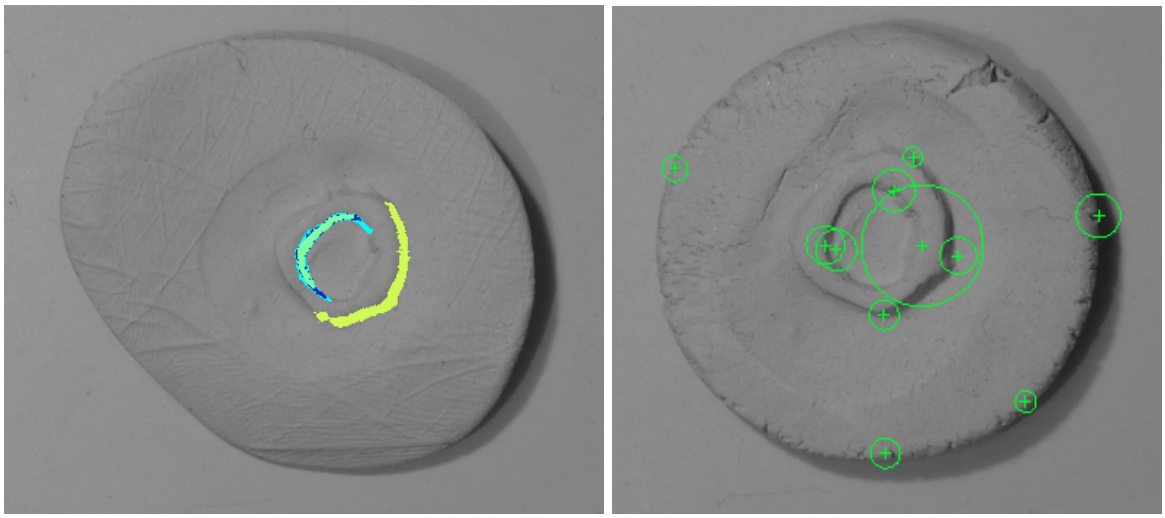

Figure 10: Comparison of MSER and SURF Characteristic Feature Responses

Figure 10 shows the characteristic feature response of two images, one which was classified with MSER features, the other with SURF. Note that the MSER responses seem to include fewer features in general, and particularly, along the border of the impression. We were abot to tune the MSER detector to largely exclude features along the sherd edge. While this intuitively seems like a better way to match impressions of the same shape, the accuracy and confusion matrices of both systems seem similar enough that it's hard to make any general objective conconclusions about their relative effectiveness.

5 Conclusions

Working on this project gave us a more complete understanding of some of the standard techniques and challenges involved in object recognition. In particular, this project emphasized the versatility of feature descriptors, but also emphasized the power of efficient and modular implementations of these tools when

first working on a project. We also gained an appreciation for some of the issues that arise working with real world data. First, hand-crafting a small labeled training set can be an effective first step to a full system analyzing a large, complex dataset. Second, spurious matches provided a reminder that effective ways to manage noise are key to the success of object recognition systems (and learning systems in general). The principles we learned on this project will guide the decisions we make moving forward with this project.

5.1 Archaeological Context and Future Work

This work represents a meaningful step towards automatic impression classification. More broadly, we have demonstrated reasonably effective application of current computer vision techniques for object recognition using widely-available tools to the classification of surface details on ceramic fragments. While analysis of artifacts in the field presents a significantly more challenging scenario, the suitability of SURF and MSER feature descriptors given controlled photographic conditions has been demonstrated.

It is important to mention that we created a custom dataset to reduce project scope to a single academic quarter, not because the problems we avoided are insurmountable. We have considered approaches for each of the difficulties we described we avoideded. For example consistent lighting is readily attaiable from existing Chavín sherd data. In particular, we have access to thousands of 3D point clouds, and we thus are able to obtain consistency by digitally synthesizing the same lighting setup for each model.

Issues of segmentation by sherd and impression are a test for image segmentation algorithms; our sense is that existing technologies are quite sophisticated and we're excited to see how techniques such as watershed segmentation handle sherd images. If we are able to segment, we see the problem of multiple impressions per sherd as a boon, as the cooccurance of two impressions on the same sherd (or collection of sherds from the same pot) is useful input to our classification system.

While we will never know absolute ground truth for many of our data points, we feel this is a tractable issue. Techniques on unsupervised learning are sophisticated and numerous. Further, we have still other ways to create a close approximation of ground truth; the trained eye could surely do an effective job at annotating a training set from images of artifacts.

Thus, we feel the pipeline we have established here is a useful backbone from which to develop a full system for Chavín sherd classification.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV*, 2006.

[2] Fatih Calakli, Ali O. Ulusoy, Maria I. Restrepo, Gabriel Taubin, and Joseph L. Mundy. High Resolution Surface Reconstruction from Multi-view Aerial Imagery. *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 25–32, October 2012.

[3] Eben Gay, David Cooper, Benjamin Kimia, Gabriel Taubin, Daniel Cabrini, Suman Karumuri, Will Doutre, Shubao Liu, Katarina Galor, Donald Sanders, and Andrew Willis. REVEAL intermediate report. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 1–6, June 2010.

[4] Benjamin Kimia and H. Can Aras. HINDSITE: A user-interactive framework for fragment assembly. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 62–69, June 2010.

[5] David Koller, Jennifer Trimble, Tina Najbjerg, Natasha Gelfand, and Marc Levoy. Fragments of the City : Stanfords Digital Forma Urbis Romae Project. *Proceeding of the Third Williams Symposium on Classical Architecture*, (Rome 1960), 2005.

[6] Lowe. Demo Software: SIFT Keypoint Detector, 2005.

[7] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[8] L J P Van Der Maaten and P J Boon. COIN-O-MATIC : A fast system for reliable coin classification M USCLE CIS benchmark. *Proc. of the Muscle CIS Coin Competition Workshop*, 2006.

[9] LJ Van Der Maaten and EO Postma. Towards automatic coin classification. *Proc. of the EVA-Vienna*, 2006.

[10] LJP Van Der Maaten. Computer vision and machine learning for archaeology. *Proceedings of Computer Applications and Quantitative Methods in Archaeology*, pages 1–7, 2006.

[11] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. *Procedings of the British Machine Vision Conference 2002*, pages 36.1–36.10, 2002.

[12] Jonah C. McBride and Benjamin B. Kimia. Archaeological Fragment Reconstruction Using Curve-Matching. *2003 Conference on Computer Vision and Pattern Recognition Workshop*, pages 3–3, June 2003.

[13] Kris Nuyts, JP Kruth, and Bert Lauwers. From a conservationist's point of view. *Proc. Conf. Optical 3-D ...*, 2001.

[14] O. C. Ozcanli and B. B. Kimia. Generic Object Recognition via Shock Patch Fragments. *Procedings of the British Machine Vision Conference 2007*, pages 104.1–104.10, 2007.

[15] Marc Pollefeys and LV Gool. 3D recording for archaeological fieldwork. *Computer Graphics and Pattern Recognition*, (June):2–9, 2003.

[16] Luciano Silva, ORP Bellon, and KL Boyer. Computer vision and graphics for heritage preservation and digital archaeology. *Revista de Informática Teórica e Aplicada*, 2004.

[17] G. Verhoeven, M. Doneus, Ch. Briese, and F. Vermeulen. Mapping by matching: a computer vision-based approach to fast and accurate georeferencing of archaeological aerial photographs. *Journal of Archaeological Science*, 39(7):2060–2070, July 2012.

[18] Geert Verhoeven. Taking Computer Vision Aloft Archaeological Three-dimensional Reconstructions from Aerial Photographs with PhotoScan. 73(January):67–73, 2011.

[19] Andrew Willis, Xavier Orriols, and DB Cooper. Accurately estimating sherd 3D surface geometry with application to pot reconstruction. *Computer Vision and Pattern Recognition*, pages 0–6, 2003.

[20] A.R. Willis and D.B. Cooper. Bayesian assembly of 3D axially symmetric shapes from fragments. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 1:82–89, 2004.