# Statistical Shape Descriptors
# for Ancient Maya Hieroglyphs Analysis

Thèse n. 1234 2012
présenté le 28 Novembre 2012
à la Faculté des Sciences et Techniques de l'Ingénieur
laboratoire l'Idiap
programme doctoral en Génie Électrique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Edgar F. ROMAN-RANGEL

acceptée sur proposition du jury:

Prof. Sabine Susstrunk, président du jury
Dr. Daniel Gatica-Perez, directeur de thèse
Dr. Jean-Marc Odobez, co-directeur de thèse
Prof. Jean-Philippe Thiran, rapporteur
Prof. Stéphane Marchand-Maillet, rapporteur
Dr. Changhu Wang, rapporteur

Lausanne, EPFL, 2012

Then I applied myself to the understanding of wisdom,
and also of madness and folly,
but I learned that this, too, is a chasing after the wind.
— Ecclesiastes 1:17

# Acknowledgements

Now I believe what I was told before: that writing down the proper acknowledgements is one of the most difficult parts of the thesis. In fact, there are many people around whose support has been of crucial importance in so many ways during the last four and a half years. Furthermore, their support often spanned over several different aspects of my life. Yet, I feel delighted to try giving the acknowledge each of you deserve, as this dissertation would not have been possible without your advises, friendship, and support. So, please allow me this attempt.

First and foremost I want to thank Fercho, Gabo, Don Tommy, and la Dra. Carmelita. You are a great family, always there to help and to encourage when things seem to go wrong. You have plenty of love to share, and you inspired many of my descisions. Special thanks to Gabo and Fercho as you make me feel welcome every single time I go back to Mexico, and you always keep saying how much you miss me, just the same I do miss you. Definitely this work could not have been done without any of you.

Then, I also want to thank my supervisor Daniel. Thank you Daniel for trusting me and for giving me the opportunity to join the social computing group to work in such an interesting topic. Thank you for all your help and your advises towards finishing this thesis, and thank you for acting not only as my advisor but also as mentor during my time at Idiap. Also, I want to thank Jean-Marc for being my co-advisor in this thesis, thank you for all your comments and advises regarding the research topic, I appreciate very much the careful observations you always come up with when reviewing a paper. I learned a lot from you both.

I want to thank the the support of the Swiss NSF through the CODICES project (grant 200021–116702), the help of the Idiap Research Institute as hosting and funding lab, and the PASCAL network for the student travel funding. Also, I want to thank Carlos Pallan, the National Institute of Anthropology and History of Mexico (INAH, Instituto Nacional de Antropología e Historia) for their valuable work as partners of the CODICES project.

I want to thank the member of my thesis committee: Prof. Sabine Susstrunk, Prof. Jean-Philippe Thiran, Prof. Stéphane Marchand-Maillet, and Dr. Changhu Wang. Thank you all for accepting being part of the committee and for all your valuable comments regarding this dissertation that helped improving it. And big thanks to Laurent and Riwal who corrected the French version of my abstract, which in practice means they almost rewrote it.

Being part of the social computing group at Idiap allowed me to meet very interesting people and making good friends. Thank you Dayra, Joan-Isaac, Laurent, Hari, Oya, Radu, Gokul, Hayley, Bogdan, Dinesh, Minh-Tri, Alvaro, Kate, Aleksandra, Eric, Darshan, Lucia, Raul, Filiberto,

for all your relevant presentations and observations during our weekly meetings, and also for the good moments and laughs during lunch and beer or coffee breaks.
Idiapers and people living in Martigny, thank you all for the good times, and specially for the Fridays' beers that helped a lot to keep on the interesting discussions or simply to get rid of the stress. Thank you Dayra, Hari, Marco, Serena, Nico, Riwal, Gwénolé, Thomas, André, Hugo, Cosmin, Francesco, Ferran, Deepu, Petr, Ivana, Leo, Tatiana, Elisa, Radu, Chris, Roy, Tiffany, Stefan, Phil, Sofi, Gari, Kenneth, Javier, Marc, Vicky, Raphael.
I would like to thank people in the fourth floor of Idiap whose work and casual chats made my time pleasant, and specially to Nadine, François Foglia, Sylvie, Valerie, Tristan, Vincent, Norbert, and Bastien.
Also, big thanks to my friends living in Lausanne that were always ready to ask "how was your week?", to help and give important support when I needed it, and to propose interesting social and cultural activities that helped distract my mind from work. Thank you Joan-Isaac, Geras, Aleja, Ganga, Robert, Andrea, Oriol, Gil, Lety, Laura, Jessi, Jose, Perrine, Aline, Bernat, Amiel, Susana. You all make of Lausanne a great place to live in.
Another group of people I want to thanks is that of my former friends back in Cuernavaca (Mexico). Thank you Laura, Ili, Omar, Doryan, Luz, Daniel, Edu, Bere, Atz, David Romero, for your friendship and for all you advises, I keep very good memories of you. Also, Laurita, Neibi, Elo, Sally, Adriana, Aline, Yiss, I am so glad we became family, friends, or simply we met again after some time. You cheered me a lot with your good mood and with the passion you put in your own life.

No doubt, each of you have largely contributed to this work, and I say thank you all.

*Lausanne, November 28$^{th}$, 2012* Edgar Roman-Rangel

# Abstract

The preservation, analysis, and management of valuable and fragile historical and cultural materials with digital technologies is a field of multidisciplinary nature, and significant to the society at large. The benefits of using visual and multimedia analysis techniques in this domain are manifold. First, automatic and semi-automatic content-based analysis methods can provide scholars in the humanities (historians, anthropologists, archaeologists, and linguists) and the arts (curators, art historians, and photographers) with tools to facilitate some of their daily work, e.g., consulting, organizing, annotating, and cataloging pieces. Second, these techniques can help obtain new insights about specific theories in the archaeological field through the recognition and discovery of patterns and connections within and across pieces in a collection. Third, automated analysis techniques can boost the creation of educational systems for public access and retrieval of digital versions of ancient media. Furthermore, the careful and efficient use of these digital collections, with potential impact in local development and tourism, also has a definite economic value.

This dissertation presents an interdisciplinary approach between computer vision and archaeology towards automatic visual analysis of ancient Maya media, more specifically of hieroglyphs. The ancient Maya civilization has been regarded as one of the major cultural developments that took place in the New World, as reflected by their impressive achievements, encompassing the artistic, architectural, astronomical, and agricultural realms. Paramount among these is their refinement of a fully-phonetic writing system that ranks among the most visually sophisticated ever created in world history. Therefore, our work is guided by realistic needs of archaeologists and scholars who critically need support for search and retrieval tasks in large Maya imagery collections. More precisely, we address the problems of statistical shape description and Content-Based Image Retrieval of Maya hieroglyphs. The type of data we analyze is rich in visual information and exhibits high degrees of visual complexity. Furthermore, the elements of the ancient Maya writing system often present inter-class visual similarity.

In this dissertation, we first present an evaluation of state-of-the-art shape descriptors for the task of shape-based image retrieval. We then introduce the Histogram-of-Orientations Shape-Context, which is a new shape descriptor designed to overcome certain drawbacks found in state-of-the-art methods, and we demonstrate its potential to deal with shapes originated from different data sources. Moreover, we present the results of using shape descriptors towards the statistical analysis of the visual evolution of Maya hieroglyphs over time and across regions of the ancient Maya world, as well as for the statistical analysis of the intra-class and inter-class visual variability of syllabic classes of Maya hieroglyphs. We also compare the performance of clustering and sparse coding techniques in the construction of efficient representations of shapes. Finally, we present results on detection of segmented symbols in large inscription based on shape descriptions.

**Keywords:** content-based image retrieval, shape descriptor, histogram of orientations, clustering, sparse coding, image detection, cultural heritage, Maya civilization, hieroglyphs.

# Résumé

La conservation, l'analyse et la gestion du fragile et inestimable patrimoine historique et culturel par le biais de technologies numériques est un domaine multidisciplinaire par nature et est importante pour la société dans son ensemble. L'utilisation de techniques d'analyse d'images et de contenus multimédias comporte de nombreux avantages. Premièrement, les méthodes automatiques et semi-automatiques d'analyses du contenu peuvent fournir aux spécialistes en sciences humaines (historiens, anthropologues, archéoloques, linguistes) et en beaux-arts (conservateurs, historiens de l'art, photographes) des outils pouvant faciliter leur travail quotidien, comme par exemple la consultation, l'organisation, l'annotation, ou l'archivage de pièces. Deuxièmement, grâce à la reconnaissance et la découverte de motifs au sein de pièces et de collections, ces techniques peuvent aider à améliorer la compréhension de théories spécifiques, dans le domaine de l'archéologie notamment. Troisièmement, ces techniques d'analyse automatiques ont le potentiel de stimuler la création de systèmes éducatifs d'accès et de récupération publics de versions numériques d'oeuvres anciennes, autrement difficilement disponibles. Enfin, l'utilisation prudente mais efficace de ces collections numériques, qui ont le potentiel d'avoir un impact dans le développement local et le tourisme, ont aussi une valeur économique certaine.

Cette thèse présente une approche interdisciplinaire entre la vision par ordinateur et l'archéologie ; ce travail ayant pour but l'analyse visuelle et automatique de collections d'images archéologiques, les hiéroglyphes Mayas en particulier. L'ancienne civilisation Maya est considérée comme l'un des développements culturels majeurs du Nouveau-Monde, comme en témoignent leurs réalisations impressionnantes, comprenant entre autres les beaux-arts, l'architecture, l'astronomie et l'agricuture. Parmi ces réalisations, le raffinement de leur système d'écriture entièrement phonétique se classe parmi les plus sophistiqués visuellement jamais créés dans l'histoire de l'humanité. Notre travail est guidé par le besoin réel, provenant des archéologues et chercheurs, d'outils de recherche et de récupération de données dans de grandes collections d'images Maya. Plus précisément, nous abordons les problèmes suivants : la description statistique de formes et la récupération d'images de hiéroglyphes Mayas. Le type de données que nous analysons est riche en caractéristiques visuelles et présentent des degrés élevés de complexité visuelle. En outre, les éléments du système antique d'écriture Maya présentent souvent une similarité visuelle entre différentes classes.

Dans cette thèse, nous présentons d'abord une évaluation de l'état de l'art quant aux descripteurs de formes pour la récupération automatique d'images. Nous introduisons ensuite l'*Histogram-of-Orientations Shape-Context*, qui est un nouveau descripteur de formes conçu dans le but de surmonter certains inconvénients liés aux méthodes existantes ; nous démontrons sa capacité à gérer des formes provenant de différentes sources de données. De plus, nous présentons les résultats de l'utilisation de ce nouveau descripteur de formes appliqué à l'analyse statistique de l'évolution visuelle des hiéroglyphes Maya dans le temps et à travers les différentes régions du monde Maya antique, ainsi que pour l'analyse statistique de la variabilité visuelle intra-classe et inter-classe de classes syllabiques de hiéroglyphes Maya. Aussi, nous comparons les performances de techniques d'agglomération (*clustering*) et de codage épars (*sparse coding*) dans la construction de représentations efficaces de formes. Enfin, nous présentons des résultats sur la détection de symboles segmentés dans de grandes inscriptions, basée sur le descripteur de formes.

**Mots-clés :** récupération d'images basée sur le contenu, descripteurs de formes, histogramme d'orientations, agglomération (*clustering*), codage épars (*sparse coding*), détection d'images, patrimoine culturel, civilisation Maya, hiéroglyphes.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Content-Based Image Retrieval (CBIR) refers to the task of searching for images inside indexed datasets using only the visual information, where computer vision techniques are often used for image description. After comparing the images in the dataset with a given query image, a subset of images considered visually similar are sorted based on a similarity score and presented as result [Smeulders et al., 2000, Datta et al., 2008].

The CBIR paradigm has been largely investigated during more than two decades, in which important improvements have been proposed for both the statistical description of images and their efficient indexing for fast retrieval applications. In the context of image description, the most common used techniques rely on the computation of intensity changes either over the whole image (*global descriptors*) or in specific regions (*local descriptors*). On one hand, the global descriptors contain statistics regarding the image as a whole, such that no further indexing steps are required. However, usually they are not robust enough to deal with variations like rotation or affine transformations. On the other hand, local descriptors only provide with information in the neighborhood of specific located regions, thus making easier the incorporation of rotation and affine invariance, although this kind of description requires the use of adequate indexing techniques to integrate the individual local descriptions into a single one [Baeza-Yates and Ribeiro-Neto, 1999, Szeliski, 2010].

Whereas the description and retrieval of gray-scale and color images has a long tradition, these tasks have probably not been investigated as much for shape binary images, where the assumptions and techniques related to local changes in the intensity values would not longer apply. Although several previous works have attempted to describe intensity images based on shape information extracted from object contours that are automatically detected [Mikolajczyk and Schmid, 2004, Lowe, 2004, Datta et al., 2008, Ferrari et al., 2008, Heitz et al., 2009], only a few works have made relevant contributions to the task of shape description [Del Bimbo and Pala, 1997, Belongie et al., 2002, Zhang and Lu, 2004, Mori et al., 2005, Yang et al., 2008, Bai et al., 2010]. However, most of those works have been evaluated on datasets containing convex silhouettes with no much of internal details [Bai et al., 2010], or on datasets with classes with little inter-class similarity [Belongie et al., 2002].

In this thesis we address the problems of shape description and retrieval of complex images, rich in visual information and in internal details, often with no convex contours, and with high degree of inter-class visual similarity. More precisely, we investigated the **statistical description of syllabic Maya hieroglyphs**, evaluating the performance of several state-of-the-art methods [Belongie et al., 2002, Mori et al., 2005] in the task of image retrieval, and proposing improvements to them. **The main contribution of this thesis is the design, development, and assessment of a new shape descriptor**. Our method has been designed to better describe the visual complexity of the Maya syllables, this is achieved by taking into account statistical information regarding the local orientations of a holistic representations of the contours of the shapes. This method has shown to achieve better retrieval results compared with previous approaches.

Our research was conducted under the context of the CODICES project, which was funded by the Swiss National Science Foundation. This project aims to develop computational tools for real needs of Maya archaeologist; and more specifically, to boost the state-of-the-art in the field of Computer Vision with the purpose of improving the automatic management of visual galleries of Maya hieroglyphs, improving common archaeological tasks such as the analysis of glyphs based on visual similarity towards their further identification, the retrieval of instances that are relevant to a visual query in terms visual similarity, and the detection of occurrences of certain glyphs in larger inscriptions.

## 1.1 Motivation

The invention of writing was a rare event in world's history, only witnessed in five different regions: in Egypt, the Indus valley, Mesopotamia, China, and the Americas. In one way or another, all other writing systems derive from these regions. Therefore, the rescue, analysis, and study of scribal material with archaeological value is of crucial importance to better understand history, and social evolution.

Maya hieroglyphs have been studied by western scholars for over two centuries, and today the field of Maya iconographic and hieroglyphic (epigraphic) analysis remains very active worldwide, given the complexity of the Maya writing system and the high frequency rate with which new archaeological sites and monuments continue to be discovered, which in turn increases the availability of source material for the research community devoted to its study, thus providing additional depth and richness to our knowledge of this ancient culture.

The ability to collect large amounts of visual material with historical value has risen the need to have computational tools for the efficient management of the resulting datasets. This is the case of the AJIMAYA project at the National Institute of Anthropology and History Institute of Mexico (INAH), whose goal is to compile a rich photographic collection of inscriptions from the ancient Maya sites that exist in the current Mexican territory. With the goal of developing the required computational tools to facilitate the work of archaeologists, Idiap started a collaboration with INAH, in which INAH helped construct a dataset of segmented syllabic instances of Maya hieroglyphs as manual drawings. These instances have been segmented from photographic material collected by the AJIMAYA project.

The decision to use this type of binary images is based on the fact that the visual format commonly used by archaeologists to study Maya hieroglyphs consists in shape representations (hand drawings) of the inscriptions.

## 1.2 Summary of contributions

The contributions of this dissertation are:

- We established contact with two word-class archaeology groups focused on the cataloging and transcription of Maya hieroglyphs: the National Institute of Anthropology and History of Mexico (INAH), and the Department of Anthropology of the Americas at the University of Bonn. These contacts matured into a project where Archaeology and Computer Science met to investigate the potential of statistical Computer Vision methods to address archaeological problems through automatic visual description and retrieval of Maya hieroglyphs.

- We collected two datasets of Maya syllabic hieroglyphs to be used for description and retrieval experiments. The collection was carried out under a constant interaction with the archaeology partners, and through an iterative process consisting in: identification, segmentation, annotation, and validation. This iterative process was repeated at different stages and spanned most of the duration of our research. This dataset represents the most comprehensive collection of Maya hieroglyphs ever analyzed with computational methods.

- We evaluated the performance of two popular shape descriptors, namely, the Shape Context descriptor [Belongie et al., 2002], and the Generalized Shape Context [Mori et al., 2005]. During this evaluation we proposed a novel approach for the comparison of complex shapes using the Shape Context descriptor [Roman-Rangel et al., 2009].

- The main computational contribution of this thesis is the design and development of a new shape descriptor, termed Histogram-of-Orientations Shape Context (HOOSC) [Roman-Rangel et al., 2011b,a]. This descriptor combines the underlying formulation of the Shape Context with the benefits that the Histogram of Oriented Gradients method provides [Dalal and Triggs, 2005], which resulted in more effective descriptors as shown by the results of a comprehensive series of retrieval experiments on Maya hieroglyphic data.

- We conducted an assessment of Sparse Coding techniques [Olshausen and Field, 1996, Lee et al., 2007] for the construction of efficient index structures of complex shapes [Roman-Rangel et al., 2012]. More specifically, we compared the retrieval performance achieved when the syllabic Maya instances are described with HOOSC vectors, and indexed as bag-of-visual-words (*bov*) constructed with both the traditional *k*-means clustering algorithm and the K-SVD algorithm (sparse linear decomposition) [Aharon et al., 2006].

- We demonstrated the generalization properties of the HOOSC descriptor by evaluating its retrieval performance on two additional dataset, one of ancient Chinese characters, and the other composed of instances of generic, standard MPEG-7 shapes. The results showed the potential to use the new descriptor for different types of binary images.
- We performed a preliminary evaluation of an approach for detection of single syllabic Maya instances in large inscriptions. During this evaluation we made use of several interest point detectors, and of two shape descriptors, i.e., SIFT and HOOSC. We discovered that the use of corners as interest points of shapes provide with better detection rates than the blob structures, namely the combination of Harris-Laplace corner detectors with HOOSC descriptors.

## 1.3 Thesis outline

This thesis is organized as follows:

- **Chapter 2** gives an overview of the Maya civilization and its complex writing system. It also introduces the archaeological needs that motivated this doctoral work, and the impact that Computer Vision technologies could have in the archaeological field. It also introduces two existing catalogs of Maya hieroglyphs.
- **Chapter 3** presents the related work in Content-Based Image Retrieval, Shape Description, Sparse Coding techniques, detection of Interest Points and their Characteristic Scale, and Shape-Based Image Detection. It also presents relevant works that have applied computer vision techniques to face cultural heritage problems.
- **Chapter 4** reviews the Shape Context (SC) descriptor and the matching framework it utilizes to rank and retrieve similar shapes. In this chapter we introduce improvements that we proposed to improve the point-to-point matching of shapes and the similarity score used for ranking of shapes. The chapter presents retrieval experiments, showing that our methodology leads to the improvement of the retrieval precision.
- **Chapter 5** reviews a variant of SC named Generalized Shape Context (GSC), and introduces the Histogram-of-Orientations Shape Context (HOOSC). This chapter explains in detail the process to construct the HOOSC descriptor, and discusses its advantages over the original SC and the GSC. It also comments its limitations. We compare its retrieval performance with the GSC on a syllabic dataset of 1200+ instances. This chapter also presents a preliminary analysis of several visual features of Maya hieroglyphs based

on statistical descriptions, such as similarity across temporal periods or regions of the ancient Maya territory.

- **Chapter 6** introduces the general sparse coding methodology, and more specifically the K-SVD algorithm which was developed as a generalization of the *k*-means clustering technique. We made use of both approaches in the task of modeling complex shapes as bag-of-visual-words, and compared them in terms of retrieval precision.

- **Chapter 7** presents an evaluation of the HOOSC descriptor on shape images from different nature to the Maya hieroglyphs. The results show the potential of this descriptor to handle several types of shape information. This chapter also introduces the first version of a visual retrieval system that was devised with the purpose of supporting archaeologist in real-world retrieval tasks of Maya hieroglyphs.

- **Chapter 8** presents the approach we followed for detection experiments given the current limitations in terms of available data. It explains the interest point detectors and shape descriptors we used, and then presents the results achieved with our formulations.

- **Chapter 9** provides a general discussion regarding the achievements of our research work, as well as its current limitations and the future work.

## 1.4 List of publications

This section lists the publications that resulted of our research in reverse chronological order.

**Journals**

- [Roman-Rangel et al., 2012]: Edgar Roman-Rangel, Jean-Marc Odobez, and Daniel Gatica-Perez. *Assessing Sparse Coding Methods for Contextual Shape Indexing of Maya Hieroglyphs.* Journal of Multimedia, Special Issue in Recent Achievements in Multimedia for Cultural Heritage. 7(2):179–192. April, 2012.

- [Roman-Rangel et al., 2011b]: Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. *Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptors.* International Journal of Computer Vision (IJCV), Special Issue in Cultural Heritage and Art Preservation. 94(1):101–117. August, 2011. Idiap Student Paper Award 2010. Best Poster Award at ENS/INRIA Visual Recognition and Machine Learning Summer School in Paris 2011.

**Conferences and workshops**

- [Roman-Rangel et al., 2011a]: Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. *Searching the Past: An Improved Shape Descriptor to Retrieve*

*Maya Hieroglyphs.* In Proceedings of the ACM International Conference in Multimedia (ACM-MM). Scottsdale, USA. November, 2011. (full paper).

- [Gatica-Perez et al., 2011]: Daniel Gatica-Perez, Edgar Roman-Rangel, Jean-Marc Odobez, and Carlos Pallan. *New world, New Worlds: Visual Analysis of Pre-Columbian Pictorial Collections.* In Proceedings of the International Workshop on Multimedia for Cultural Heritage (MM4CH). Modena, Italy. April, 2011.

- [Roman-Rangel et al., 2009]: Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. *Retrieving Ancient Maya Glyphs with Shape Context.* In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Workshop on eHeritage and Digital Art Preservation. Kyoto, Japan. September, 2009.

# 2 The ancient Maya writing system

The pictorial material we have analyzed is not a modern construct, it was devised several hundred years ago by a now extinct civilization: the ancient Maya culture. In this chapter we provide an introduction to the ancient Maya culture, and more specifically, to the Maya writing system. We also stress the importance and motivation to conduct research in the field of computer vision with archaeological applications.

## 2.1 The Maya World

The Maya is regarded as the epitome of the ancient (pre-industrial) civilizations in the Americas as reflected by their impressive achievements, which encompass the artistic, mathematical, architectural, astronomical, and agricultural fields, many of which are comparable to those of the Old-World cultures that developed in Egypt, Greece, Rome, Sumer, and Babylon. Paramount among these, is their refinement of a fully phonetic writing system which ranks among the most visually sophisticated ever created in world history [Stuart et al., 2005].

Roughly outlined, the ancient Maya was one of several civilizations belonging to a cultural super-area called Mesoamerica, which encompassed the major parts of what are now the countries of Mexico, Guatemala, Honduras, Belize, and El Salvador. Figure 2.1 shows the area that was occupied by the Maya culture. The Maya culture began to flourish during a chronological period called the Pre-classic (c.a., BC 2000 - 250 AD). Although their development was differential according to region and speed, generally speaking, their heyday is regarded to have occurred during the subsequent Classic period (c.a., AD 250 - 900), and it was then when their hieroglyphic writing and the highly encoded iconographic imagery attained the levels of sophistication and consistency that we can rightly regard as a coherent, self-contained visual system, capable of conveying speech and ideas with admirable precision, even when compared with our current devices for information exchange: e.g., alphabets, syllabaries, graphic conventions, and so forth.

Maya writing was not an isolated phenomenon, but stems from a larger phonetic tradition

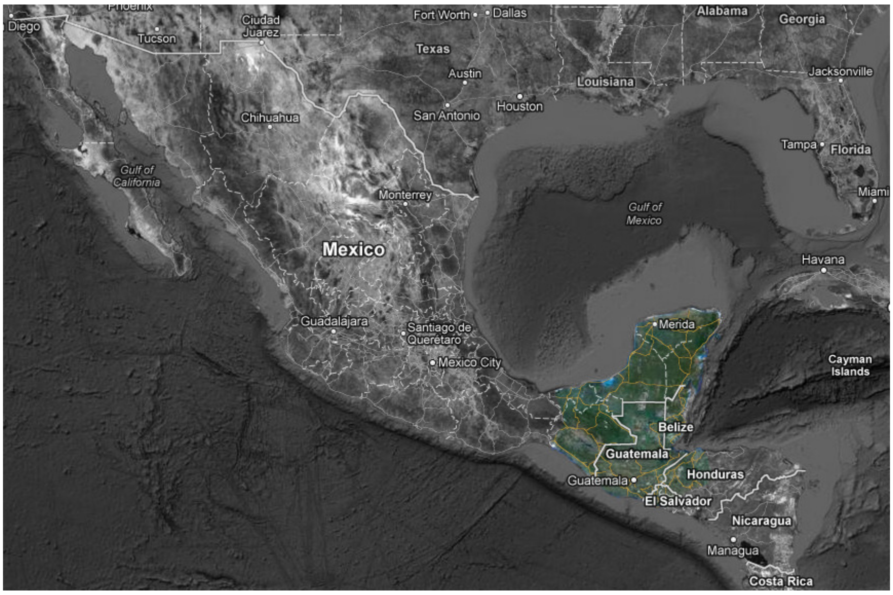


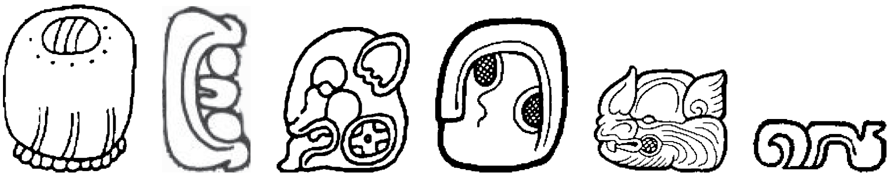
Figure 2.1: Maya region highlighted in color. Map edited from Google Maps ©.

that developed in southern Mesoamerica [Justeson et al., 1985, Stuart et al., 2005]. Some of the earliest Maya inscriptions date from the late Pre-classic (c.a., 400 BC - 250 AD), and originated in the Guatemalan Petén district, the Salama Valley, and the Pacific coastal region, and at sites such as San Bartolo, Takalik Abaj and El Baul. Later, it spread into the western and northern lowlands of the Usumacinta, Campeche, Belize, Quintana Roo, and Yucatán, remaining operational during at least 17 or 18 centuries. By late-Classic times (c.a., 600 - 900 AD), the usage of the Maya script became commonplace throughout the entire Maya lowlands, and has been detected on a few but steadily growing number of transitional lowland-highland sites as well. During the Terminal Classic (c.a., 800 - 950 AD), writing and scribal traditions continued, albeit on a fairly diminished capacity, with the exception of new and revitalized northern centers like Chichen Itza, Uxmal, Mayapan, and Santa Rita Corozal, all of which continued to operate after the so-called “Maya Collapse” (c.a., 950 AD) [Sharer, 1996].

## 2.2 The Maya Writing System

In a nutshell, any ancient script could be defined as a system for the visual recording of information through signs (graphemes) related in some way to the meanings (lexemes) and sounds (phonemes) that conform any given speech [Browder, 2005].

The Maya script belongs to the class of the logo-syllabic writing systems, to which a large number of other ancient-world scripts belong, such as the Anatolian from Syria, or the Hiragana from Japan. This term describes systems composed of two functionally distinct categories



(a) *b'a* (b) *u* (c) B'AHLAM (d) KAB' (e) SUUTZ' (f) K'AHK'

Figure 2.2: Segmented Maya syllabographs (a - b) and logographs (c - f). ©AJIMAYA.

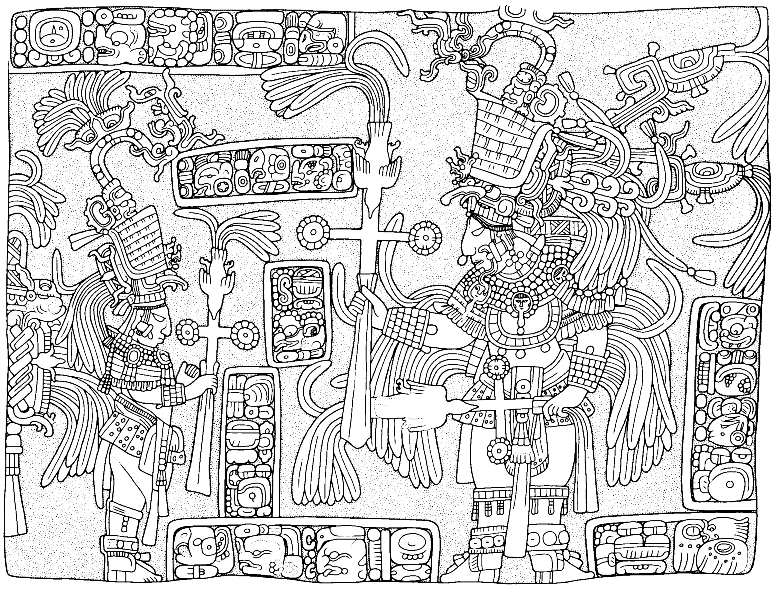


Figure 2.3: Maya inscription from Yaxchilan site. ©AJIMAYA.

of signs: syllabographs and logographs [Lacadena, 1995]. The former are visual signs which encode only specific phonetic value (i.e., phonemes) and almost always comprise a consonant-vowel or a single aspirated vowel structure, (denoted respectively with lower case cv or v) e.g., Maya syllables *b'a* and *u*. On the other hand, logographs encode both sound and meaning, being a rough equivalent to the notion of "word-signs", and the vast majority of them have a consonant-vowel-consonant structure denoted in upper case (CVC) with different possible compounds, e.g., CVC - VC, CVC - CVC, etc., for instance the Maya word B'AHLAM meaning "jaguar". Note that the embedded vowel in the logographs could be either simple or complex, thus making possible forms like KAB' (earth), SUUTZ' (bat), and K'AHK' (fire). Figure 2.2 shows examples of syllabographs and logographs.

A third type of inscription is the Maya visual narrative or iconography, which is a substrate of Maya art, and that was used to tell stories both of mythological and historical nature. Figure 2.3 shows an inscription containing syllabographs, logographs, and Maya narrative.

The common practice of the Maya script consists of combining both syllabographs and logographs. In practice, several signs of both types are arranged inside a single glyph-block,

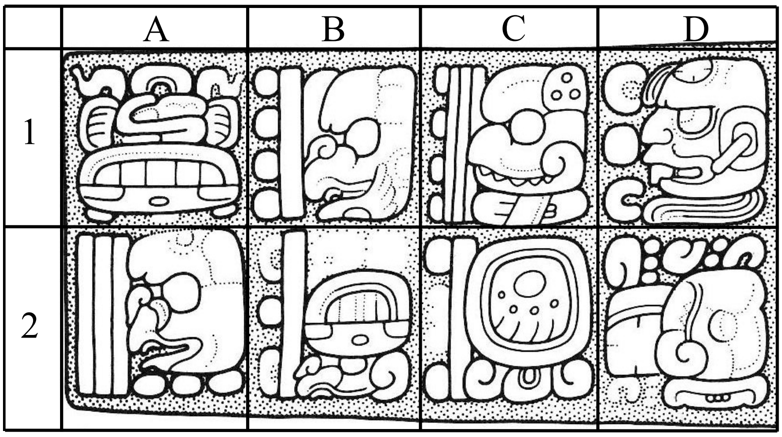


Figure 2.4: Glyph-blocks examples. ©AJIMAYA.

where usually logographs are phonetically complemented by syllables, either on initial position (i.e,. as prefix or superfix) or in final position (i.e., as postfix or suffix). In turn, the glyph-blocks are arranged on a paired columnar format, which is referred to as a system of coordinates, where letters designate the columns and numbers the rows. In such a grid, a standard reading order for a text comprising 4 columns and 2 rows would be: A1, B1, A2, B2, C1, D1, C2, D2. Figure 2.4 shows an example of such a paired column format.

One of the main characteristic of the Maya writing system is the common visual modification of certain characters. Namely, several phenomena can be found:

(a) *Infixation*, which involves one sign being reduced in size and inserted within another, as the example (a) in Figure 2.5, where one of the sign on the top-left of the image is resized and then inserted inside the other one, the result of this insertion is shown in the bottom-left of the image.

(b) *Conflation*, which occurs when two signs are visually fused, each retaining its same relative size. For instance, the case (b) in Figure 2.5 corresponds to the conflation of the two glyphs on the top-left of the image. Note that certain visual features suggest the fusion of the two signs, such as the round pattern in the bottom that contains a small circle, or the double contour of the external boundary.

(c) *Superimposition*, that takes place when one sign partially covers another which main elements remain visible as background. The sign shown in the left side of case (c) of Figure 2.5, is partially covered by another glyph, and the superimposed representation is shown on the right side.

(d) *Pars pro toto*, which makes use of one or more diagnostic features of any given sign in order to mean the whole. For instance, the sign on the top-right of Figure 2.5 that has the visual patter of a 'face placed between two circles' corresponds to a single glyph. In case

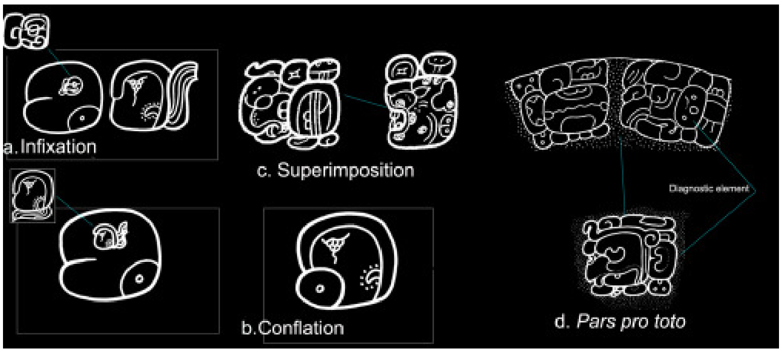


Figure 2.5: Common visual modification of Maya hieroglyphs. Source: Carlos Pallan.

(d) at the bottom-left of the image, the same sign is represented by only the central circle containing the face patter.

By taking into account both syllabographs and logographs, an approximate of 1000 distinct signs have been identified thus far, from which only a maximum of 800 were used at any given time during the ancient Maya world. Currently, only 80% to 85% of the known hieroglyphs have been deciphered and are readable.

Among the many challenges for the decipherment, as well as for the automatic processing of the Maya hieroglyphs, there is the high level of intra-class visual variability, which increments as the temporal and spatial gaps increase, and the inter-class similarity among several of the hieroglyphic classes [Grube, 1989, Lacadena, 1995]. Another key issue to face towards the decipherment of Maya hieroglyphs is the relative low rate at which certain hieroglyphs are observed in the inscriptions, either because they were only used at specific sites or at specific periods of the ancient Maya world, or simply because not enough exemplars have been discovered yet. Thus, for the purpose of creating an image dataset with the highest applicability for our research, and under the logic of a progressing scheme, we decided to focus exclusively on syllabic signs reserving logographs for future work, going from relative simplicity towards increased complexity, which ideally could lead us to also process Maya iconographic elements in the future.

## 2.3 The AJIMAYA and CODICES projects

Our source of data is the AJIMAYA project (*Acervo Jeroglífico e Iconográfico Maya* - Hieroglyphic and Iconographic Maya Heritage), which started in 2006 as an effort of the National Institute of Anthropology and History of Mexico (INAH, *Instituto Nacional de Antropología e Historia*).

AJIMAYA's goals encompass the safekeeping, preservation, study, and dissemination of monuments and written records from Mexico's Maya archaeological and cultural heritage, which number in the order of thousands, and include several UNESCO's World Heritage sites. This project has been partially inspired by the work of Ian Graham [Graham, 1975], who along with his collaborators at Harvard University's Peabody Museum of Archaeology and Ethnology, provided the scholarly community with a comprehensive repository of photographs and drawings of such cultural items.

For some of the Mexican Maya sites, the AJIMAYA project has already compiled a full photographic record of monuments. The photographic material obtained has to undergo an eight-fold methodological treatment that generates the data commonly used by archaeology scholars, and that consists in:

1. Digital photographs of inscriptions are taken in sites at night under raking-light illumination, which brings out the level of detail that facilitates the study of eroded monuments.
2. Line drawings are manually traced on top of multiple layers of enhanced photographs under different light conditions. This is done to capture the inner features that are diagnostic towards their subsequent identification.
3. Manual segmentation, search, and identification of hieroglyphs is done by consulting existing glyphic catalogs.
4. Manual transcription, performed by experts by rendering the phonetic value of each Maya sign into alphabetical conventions.
5. When needed, transliteration is performed to represent ancient Maya speech into modern alphabetic forms.
6. Morphological segmentation breaks down recorded Maya words into their minimal grammatical constituents (morphemes and lexemes).
7. Grammatical analysis uses common conventions to the fields of historical and structural linguistics to indicate the function of each segmented element.
8. Final translation renders the ancient Maya text into a modern target language, e.g., English.

Figure 2.6 shows the first and second steps of this process.

An initiative from the Idiap Research Institute led to the CODICES project, as a collaboration with INAH to conduct research in the area of computer vision with the purpose of improving the process of accessing the digital material. More specifically, in this thesis we have addressed the problem of statistical description and automatic retrieval of hieroglyphs, thus partially addressing the improvement of the third step of the generation process of digital material of Maya inscriptions.

Our collaborative research is the starting point towards future work facing more challenging issues, such as the development of a highly refined hieroglyphic catalog, having the capabilities to be periodically updated, and to incorporate input from multiple scholars working

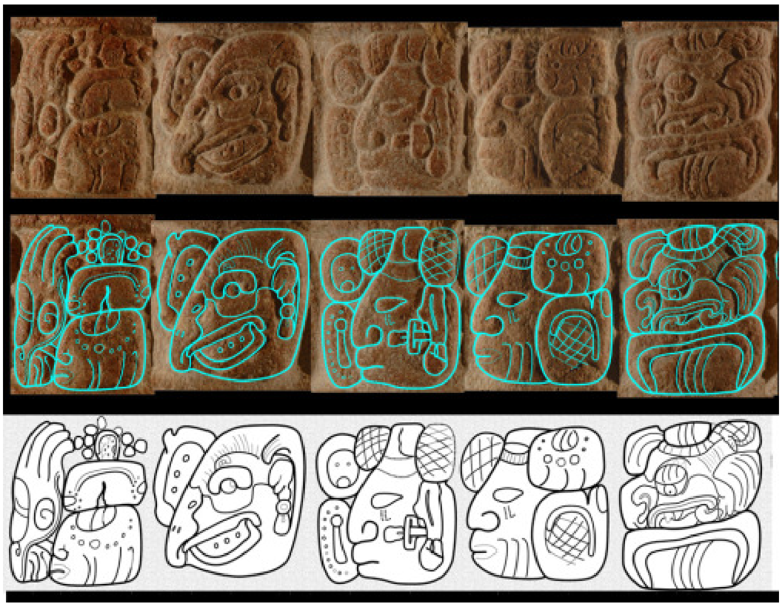


Figure 2.6: First two steps in the process of generating digital material. ©AJIMAYA.

in different parts of the world. Despite pioneering efforts on this regard [Thompson, 1962, Macri and Looper, 2003] (see section 2.4 for more details), one of the challenges that limits the potential of existing catalogs is found at the taxonomic level, as epigraphy needs an upgraded system for classifying the 800+ known signs, which separates the consensually deciphered from the undeciphered ones, where the authorship of each specific decipherment is unambiguously indicated, and where all known contexts of occurrence of each particular sign are readily accessible, in order to better assess its function within the script as a whole.

In order to succeed in creating these and other research tools, one of the basic abilities that needs to be developed is that of performing queries of a particular glyph, which could retrieve the contexts where equivalent instances occur, and also having the ability of automatically detecting not only specific instances of particular signs, but also their variants, such as their *allographs* (different signs indicating the same phonetic value), their *homophones* (signs with similar sound, which meaning could differ), and their *polivalencies* (signs which could take more than one phonetic value, depending on the context).

## 2.4 Existing catalogs

One of the earliest work on cataloging Maya hieroglyphs is [Thompson, 1962]. This work represents an important milestone in the field, as it introduced indexing numbers to the signs, making possible more efficient references such as “T229”, where “T” stands for Thompson, and the number is a consecutive identifier that allows including new symbols as they are identified, thus sometimes referred to as T-number. This system was definitively far more efficient than previous references based on description, for instance “the sign found in such position of such inscription" or "the sign with the shape of a jaguar". This catalog includes more than 1000 hieroglyphs from codices and monuments, including visual variation for some of them. The Thompson's Catalog is currently unavailable for scholarly research from the University of Oklahoma Press.

New signs have been discovered as the research advances towards deciphering the Maya writing system, and new knowledge was acquired. For instance, it was discovered that the consecutive Thompson numeration does not correspond to any kind of taxonomical ordering, and that new ways to classify the glyphs was need. "The New Catalog of Maya Hieroglyphs, Volume 1: The Classic Period Inscriptions" [Macri and Looper, 2003] (Macri&Looper) introduces a different notation based on two letters that identify each hieroglyph as: main sign, affixed, head profile of an element, and so on. This notation is complemented with a consecutive numeration that allows for expansion as new elements are discovered, e.g., MR1, MR2, MR3. This catalog also provides references to identify the glyphs by their T-number as for many years the Thompson catalog was the standard resource to identify Maya signs. The Macri&Looper catalog represents the most up to date and most comprehensive guide to the Maya symbols of the Classic period.

The website of the Foundation for the Advancement of Mesoamerican Studies, Inc. (FAMSI, www.famsi.org) [FAM], contains a vast collection of different resources for archaeological research, including catalogs, maps, journal papers, references, and guides to ancient writing systems, most of them been of public access for academic purposes. Such is the case of the booklet "Writing in Maya Glyphs" [Pitts and Matson, 2008], which has served as third source for some of the data used in this thesis.

The data used for our research consists mainly in syllabic instances of the Maya writing system, most of which were extracted from inscription of the AJIMAYA corpus. In order to increase the size of our datasets, we also included instances from the catalogs introduced in this section, i.e., [Thompson, 1962, Macri and Looper, 2003, Pitts and Matson, 2008], which in turn, increased the visual variability of our datasets.

## 2.5 Conclusions

In this chapter, we have briefly introduced the Maya culture that developed in ancient Mesoamerica and its sophisticated writing system. The Maya writing system, containing around a thousand hieroglyphs with high levels of visual detail, poses difficulties to both manual and automatic identification. Some of these hieroglyphs were operational during more that 17 centuries and spread over several territories, thus incorporating new visual features as they were disseminated.

The National Institute of Anthropology and History of Mexico (INAH) has gathered a large collection of digital photographs of many Maya sites. However, the current process of translation and interpretation of Maya inscription requires several steps of hard manual work. Thus, the development of automatic techniques for statistical visual description and semi-automatic identification are of high interest for the archaeological community.

The CODICES project represented a bidisciplinary collaboration to conduct research in the areas of Computer Vision and Content-Based Image Retrieval with the goal of improving the analysis of Maya inscriptions. Such a collaboration resulted in this dissertation.

# 3 Related work in Computer Vision

The statistical visual description and the semi-automatic retrieval of shapes (binary images) are the Computer Vision topics of main interest in this work, and particularly in the context of digital image collections with cultural and historical value. Besides those topics, we also conducted research in the efficient indexing and the detection of shapes. In the following, we discuss some related work in all those areas.

## 3.1 Description and retrieval of shapes

The use of computer vision techniques for describing, indexing and retrieval of 2-D imagery has been the topic of important research for more than two decades [Smeulders et al., 2000, Zhang and Lu, 2004, Datta et al., 2008], during which, important milestones have been reached. Yet some issues remain to be properly tackled.

The works of [Zhang and Lu, 2004, Yang et al., 2008] provide relatively recent reviews on shape representations. Roughly speaking, shape descriptors differ according to whether they are applied to contours or regions, and whether the shapes are represented globally or by the combination of local structures.

Global representations like Fourier or moment descriptors are usually sensitive to variations in some regions of the shape. The use of invariants shape moment as global shape descriptors has a long tradition [Hu, 1962, Teh and Chin, 1988, Wang et al., 2011b]. Descriptors based on moments can be relatively easy to compute, and they achieve good levels of robustness against location, scale, and rotation changes. Among the several method based on moments, Zernik moments seem to perform best [Teh and Chin, 1988]. Although these methods work well to discriminate among different classes, they perform poorly when the degree of affine transformations is high, and for complex shapes whose instances have many local variations [Wang et al., 2011b].

Fourier descriptors are an alternative for simple shapes with convex contours [Zahn and Roskies, 1972], and the original shape can be recovered by the used of the inverse Fourier transform. This techniques have achieved good results in shape classification and shape alignment of simple instances [Duan et al., 2008]. However, they introduce the need for efficient approaches to normalize descriptors derived from different shape signatures, they also require comprehensive explorations to found the appropriate number of coefficients for the Fourier descriptors [Duan et al., 2008].

Sketch-based image retrieval is topic that has attracted much attention [Del Bimbo and Pala, 1997, Banfi and Ingold, 1999, Cao et al., 2011, Eitz et al., 2012], as it could allow to retrieve both shapes and images by querying hand-drawings. One approach to this problem is the use of splines to estimate elastic deformation of user sketches, and then solve the problem as template matching [Del Bimbo and Pala, 1997]. The case of retrieving color and intensity images is particularly challenging as it requires mapping descriptors of different nature [Banfi and Ingold, 1999, Wang et al., 2011a].

The Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005] in another global descriptor proposed for human detection, and is one of the most effective descriptors for intensity images. It relies on histograms of local orientations computed inside connected regions of the image that latter are grouped into overlapping blocks. As the histograms correspond to localized segments of the image, it is difficult to adequate this approach for rotated images (usually not required for pedestrian detection).

The representation and matching of visual entities was improved largely with the use of robust local viewpoint-invariant features computed over automatically detected areas, [Mikolajczyk and Schmid, 2004, Lowe, 2004]. The local character of these features provides robustness to image clutter, partial visibility, and occlusion, while their invariant nature addresses issues related to changes in viewpoint and lighting conditions. In this context, one emerging research direction has focused on modeling objects by histograms of quantized local descriptors (bags-of-visterms, *bov*) [Sivic and Zisserman, 2003, Quelhas et al., 2005], which allows for fast retrieval applications in large collections. However, the main limitation of this approach is that the spatial information is lost. Recently, bag-of-features approaches able to retain spatial information has been proposed [Lazebnik et al., 2006, Cao et al., 2010] showing promising results in the tasks of retrieval of building images and scene recognition. However, while retaining spatial information, those approaches require the use of supervised learning methods to achieve invariance against rotation and affine transformations, and might result in high dimensional descriptors. In addition, all of the above mentioned works represent the visual content with appearance descriptors that rely largely on the intensity information of the image, which might not be well adapted to shape (binary) images.

In contrast to global descriptions, structural approaches which represent objects by trees of local segments do not suffer from the drawbacks like the alterations or variations in certain regions [Zhu et al., 2008, Lu et al., 2009]. However, tree representations become very large when dealing with complicated shapes and make these techniques computationally very expensive for shape comparison. Therefore, it results unfeasible to use most of these techniques to deal with complex hieroglyphs.

Recently, the use of shapes to represent objects regained some attention, focussing mainly on learning models from few training images [Jiang et al., 2009], shape discovery [Lee and Grauman, 2009], model matching in images [Zhu et al., 2008, Lu et al., 2009], or exemplar-dependent shape distance learning [Frome et al., 2007]. Also, shape models for descriptive models can be learned combining long salient bottom-up contours [Srinivasan et al., 2010], where a latent SVM learning formulation tunes the scores of a many-to-one contour matching approach used to deal with the random fragmentation that might occur in the contours. However, in most cases these works rely on finding feature correspondences and/or shape alignment using techniques that are too slow for retrieval applications in large databases (e.g., the thin plate spline robust point matching algorithm [Jiang et al., 2009]). In addition, due to their different focus, they mainly address the representation of intensity images based on their shapes, rather that representations of shapes of 3-D real objects. Also, currently they handle object categories with limited complexity, like apple logos, mugs or bottle shapes, which are substantially different from the complex Maya hieroglyphs.

Taking inspiration from the success of the appearance local descriptors for object representation, recent works have investigated similar approaches that construct visual vocabularies from boundary fragments to represent shape information [Shotton et al., 2005, Opelt et al., 2006, Ferrari et al., 2008]. As example, [Ferrari et al., 2008] uses quantized representations of local segment-networks as structural shape representations. However, these techniques were specifically designed for object detection or classification, and are thus associated with discriminant learning approaches and sliding-window (or voting methods for localization). Therefore they can be explored for localization of shapes, but they would not provide appropriate matching scores for retrieval applications.

In contrast, the Shape Context algorithm [Belongie et al., 2002] provides with holistic shape representations through semi-local descriptors that integrate shape and geometric discrepancy measures, providing a good framework for shape comparison. In addition, its generalized version [Mori et al., 2005] incorporates local information about the contour orientations, and uses quantized descriptors that allow for fast retrieval implementations. However, the Shape Context has two main drawbacks: the simple count of pixels does not seems to be enough for shapes with contours of varying thickness; and the methodology to estimate its characteristic scale makes the descriptors very large, thus loosing their locality. Also, using only the dominant orientation as the Generalized Shape Context does, might result in noisy signals when this method is applied to binary images. Part of our work was devoted to propose improvements to these techniques, which resulted in the development of a new descriptor.

The use of thin-shape representations as input for shape description has been also studied. In [Sundar et al., 2003], both the geometric and topological information of 3-D objects is combined in the form of a skeletal graph, using graph matching techniques to match and compare skeletons. On a different direction, the shape recognition problem is also approached with a graph matching algorithm in [Bai and Latecki, 2008], based on object silhouettes where geodesic paths between skeleton endpoints are compared without considering the topological graph structure. More recently, shape retrieval has been boosted with graph techniques like local diffusion process [Yang et al., 2009b] and graph transduction [Bai et al., 2010], achieving very good retrieval results. However, rather than dealing with complex shapes, these methods focus on retrieval of convex silhouettes with well defined boundaries, and usually without internal details.

## 3.2 Indexing of shapes

As mentioned in section 3.1, the use of histograms of local descriptors has led to fast retrieval applications in large collections [Sivic and Zisserman, 2003, Quelhas et al., 2005]. We have investigated the use of sparse coding techniques for vector quantization and the construction of *bov* representations for the retrieval of Maya hieroglyphs. Below, we present the related work in these directions.

Sparse Coding was first introduced in [Olshausen and Field, 1996, 1997] as a method to find sparse linear combinations of basis functions with the purpouse of encoding natural images in a similar way the human brain does. Given that the resulting sparse image codes have a high degree of statistical independence, the authors suggested that they are more suitable to be used for the latest stages in image processing. Even though the authors do not provide any quantitative evaluation of their method, they show that the sparse coding of natural images leads to a set of localized, oriented, bandpass fields that are similar to those found in the primary visual cortex of mammalians.

Since the seminal works in sparse coding, a large number of works have used this approach in image and video processing, multimedia indexing, and image classification [Wright et al., 2010]. For instance, based on stochastic approximations, an online optimization algorithm for dictionary learning was proposed in [Mairal et al., 2010] for in-painting and image restoration. The KSVD algorithm was introduced in [Aharon et al., 2006] as a method to estimate dictionaries with orthogonal basis, and to compute sparse representations. This method was applied for restoring facial images and for image compression. It was extended in [Mairal et al., 2008] to multi-scale sparse representations for the enhancement and restoration of color images and videos. In our work, we evaluate the applicability of the KSVD algorithm to deal with shape representations of Maya hieroglyphs. A previous work that investigated a similar problem is [Ranzato and LeCun, 2007], where the authors presented a method to extract shift-invariant sparse features of shapes. This method was used to train a deep convolutional network for classification of shape images of numeric digits, and for compression of text document images achieving state-of-the-art results. However, digit shapes are far simpler compared with the high visual complexity of the Maya hieroglyphs.

In another direction, the problems of shape representation and recognition of multiple objects in images were approached with sparse decompositions of low-level features in [Mendels et al., 2006]. However, these approaches were mainly evaluated on synthetic data, detection of simple shapes in aerial images, and reconstruction of brain magnetic resonance images in a qualitative manner. In general, there are a very few works that addressed shape encoding (rather than shape images) using sparse coding.

Since the *bov* representation is widely used in the image retrieval community [Baeza-Yates and Ribeiro-Neto, 1999], there has been great interest in the validation of sparse coding techniques for the construction of *bov* representations. One of the initial works for object matching in videos based on *bov* is [Sivic and Zisserman, 2003], where objects are represented by quantized sets of viewpoint invariant region descriptors. The use of spatial pyramid matching as a method to generalize vector quantization to sparse coding has been also investigated [Yang et al., 2009a] by the use of SIFT sparse codes for image categorization obtaining state-of-the-art performance. This work was extended in [Gao et al., 2010] by the use of a Laplacian constraint, which overcomes the loss of spatial information of the *bov* construction process. However, non of them addressed the use of sparse coding for representation of complex shapes.

Several pooling schemes of sparse coding for vector quantization were evaluated in [Boureau et al., 2010a,b], where a set of experiments for feature recognition and image classification, showed that some pooling strategies perform better than others. Besides the success of sparse coding in the representation of natural images, a recent work in image recognition [Rigamonti et al., 2011] has suggested that sparse coding might not be suitable if the input signal contain a reasonable level of noise.

## 3.3 Interest points and Scale space

The local and semi-local features we used during our research were computed following the ideas presented by [Belongie et al., 2002]. More specifically, we described points that were randomly and uniformly selected from the contours of the shapes, and the locality scope for each selected point is estimated as a function of the average pairwise distance computed among all the points in each shape. However, this approach might not be suitable when the bounding boxes of specific object are unknown, such as in the case of image segmentation or localization. Since, the detection of interest points and their characteristic scale has been largely investigated for intensity images, we comment in this section the most representative works on these matters.

Interest point at the center of blob-like structures are localized by finding maximum responses in the 3-D scale-space formed by the convolution of an intensity image with a 2-D Gaussian filter and with a Laplacian operator that recovers the intensity changes in the neighborhood of each pixel, hence the name *Laplacian of Gaussians* [Crowley and Parker, 1984, Lindeberg, 1998]. This approach has been made more efficient by the use of a scale-space pyramid built with *Difference of Gaussian*, in which the image is smoothed and sampled by Gaussian kernels at different scales, and later the subtraction of each two consecutive smoothed images is performed, such that local extrema in this space directly define interest points and their characteristic scale [Lowe, 2004].

Another kind of interest points are corner points. The *Hessian* corner detector combines the determinant and the trace of the Hessian matrix computed on local sections of the intensity image. The use of the determinant avoids detections when the second derivative indicates changes in only one direction [Mikolajczyk and Schmid, 2002]. The *Harris* detector is similar but it uses only the first derivative [Harris and Stephens, 1988, Mikolajczyk and Schmid, 2002]. In [Mikolajczyk and Schmid, 2004] a method that combines the Harris corner detector with the Laplacian scale selection is presented (*Harris-Laplace*), which is able to handle not only variation in scale and location, but also affine transformations. This method has shown improved results in terms of repeatability of the detected points, and in the task of image matching, both on datasets of intensity images.

The works of [Aanæs et al., 2012] presents a comparison of different interest point detectors for 3-D scenes of houses, building materials, fruits and vegetables, fabric, and printed media. They have found that the Harris corner detector performs best, followed by the Hessian detector, and the difference of Gaussians. However, they highlight that the choice of a interest point detector might depend on the application.

Finally, the problem of detecting interest points and their scale space for shapes was studied under the specific rule of detecting salient convex local arrangements of pixels [Jurie and Schmid, 2004]. This approach estimates the convexity of the local sections of the shape contours measuring how good they match a circle template, the it combines the convexity measure with the orientation entropy of the local contour section. The evaluation of this method for object detection showed improvements over traditional intensity-based approaches. However, this approach might not perform as good for shapes that are rich in internal details that could add noise to the convexity index, or when the shapes contain very few convex local structures.

## 3.4 Shape Detection

Several approaches have shown success in the task of detecting objects on gray-scale images [Viola and Jones, 2001, Dalal and Triggs, 2005]. The common framework for image detection implements a sliding window approach [Viola and Jones, 2001, Lampert et al., 2008], in which a classifier is used to evaluates subwindows and decides whether or not the element of interest is present on them. Often, the images are described using local descriptors such as SIFT [Lowe, 2004] or HOG [Dalal and Triggs, 2005], or by template matching approaches [Viola and Jones, 2001, Payet and Todorovic, 2011].

The problem of detecting shapes is often investigated as object detection based on shape information, i.e., it is tackled by extracting contours and local orientations based on the local gradients of intensity images. For instance, [Ferrari et al., 2008] propose the use of networks of local segment as shape descriptors, and performs detection of shapes belonging to classes that are relatively easy to differentiate in visual terms. With a similar approach, the detection of objects is improved by using deformable shape models that are built combining the same network of local segments with a Hough voting scheme [Ferrari et al., 2007]. A more recent work performs object detection and pose estimation by using view-dependent shape templates constructed based on contour segments [Payet and Todorovic, 2011].

A loop-like variation of a network of segments is proposed based on landmark points that are connected [Heitz et al., 2009]. This approach provides with a probabilistic model for shape description that helps localize objects for descriptive classification. However, the evaluation of this method was done on classes that are considerably different to one another in visual terms, and whose instances contain the object of interest rather on fixed locations.

As already stated, all of the above mentioned method perform object detection based on shapes, rather than shape detection. Also, they address this problem based on classification of sub-windows, which requires enough data for training, which might not be suitable for relatively small datasets.

## 3.5 Cultural heritage and art preservation

The use of computer vision techniques to improve the access to material with cultural value has become popular [Stanco et al., 2011], going from image description and classification [Zambanini and Kampel, 2011], registration and 3-D reconstruction of real-world structures [Agarwal et al., 2009, Larue et al., 2011], enhancement and restoration of images [Whyte et al., 2010, Russell et al., 2011], authorship identification [Li and Wang, 2004, Johnson et al., 2008], to efficient indexing and web presentations [Bagdanov et al., 2011].

A system for retrieval of paintings and photos of art objects by using content metadata is presented in [Lewis et al., 2004]. This system was developed by a multidisciplinary team jointly with a group of European museums. The project identified two relevant user tasks from specialists' feedback: *query-by-example*, and *cross-collection search*, and proposed algorithms based on adaptation and extension of techniques developed in the mid-90s (e.g., color coherence vectors). Another work on access to cultural content consist in a pipeline for web publications [Larue et al., 2012]. This work uses a 3-D scanner to automatically generate Web digital models of cultural pieces such as ceramics and vessels. The process is such that color and texture are used to enrich the geometric information. It also makes use of standard technologies for efficient storage and retrieval of the captured models.

A discussion about variations of visual query formulations in cultural heritage collections, and ideas to tackle this problem based on queries by region-of-interest approaches appears in [Boujemaa et al., 2002]. Techniques for detection of visual patterns and visual trends in image collection are used for characterization of artist styles in ancient Chinese paintings [Li and Wang, 2004]. In [Zhuang et al., 2007], Chinese calligraphy characters are retrieved using contour shapes and interactive partial-distance-map-based high-dimensional indexing that speeds up the performance. The recognition is historical documents is approached using a

OCR system that learns the patterns of characters in binary images in [Vamvakas et al., 2008].

The classification tasks for artist identification is evaluated using wavelets that characterize brushstrokes of several van Gogh paintings [Johnson et al., 2008]. This is a very interesting work as the accurate identification of authorship is of great relevance in the cultural field. With similar goals, the tasks of artistic style recognition and authentication were successfully performed with sparse models to distinguish drawings by Pieter Bruegel the Elder from its imitations [Hughes et al., 2010].

For the specific task of describing characters of ancient inscriptions there have been important contributions as well. For instance, the analysis and prediction of sequences of an undeciphered script from the Indus civilization is conducted using Markov and N-gram models [Rao, 2010], showing that bigrams are able to discover syntax (order of symbols) even in undeciphered scripts. On a different context, the work in [Frauel et al., 2006] addresses the recognition of polymorphic symbols with a set of rules such as single symmetry axis and morphology, those rules were applied to recognize one Mesoamerican symbol based on the description of sets of discrete curves.

Overall, the use of computer vision techniques as tools to access and manage material with cultural value has become common, as these techniques have proven effective to perform tasks that otherwise would be conducted manually, thus consuming much time and increasing the chance for human errors. It is expected that computer vision methods will gain even more popularity for these kind of tasks in the future.

## 3.6 Conclusions

In this chapter we discussed relevant work that has applied Computer Vision methods to facilitate the access to visual material with cultural value. Works on that multidisciplinary intersection have achieved interesting results facing problems such as image classification, 3-D reconstruction, image restoration, authorship identification, etc.

We also commented the most representative work in areas that relate to our research. More specifically, we discussed the contributions and drawbacks of methods for image description, shape retrieval, indexing of shapes, detection of interest points, characteristic scales, and detection of shapes.

The statistical description of shapes is the main problem tackled in this dissertation. Shape descriptors can be grouped in global and local descriptors. Some of the most important early methods for shapes description were based on moments or Fourier descriptors, these descriptors work well for shapes with low levels of local variations. Recently, the use of histograms of orientations has proven robust agains a variety of transformations, specially when implemented as local descriptors. The use of Shape Context formulations has achieved good retrieval results of shape instances.

We reviewed the characteristics of several shape descriptors, and identified the Shape Context as a suitable descriptor for Maya hieroglyphs. In the following chapters we present the contributions we made to the shape description based on Shape Context-like formulations. More specifically, in chapters 4 we propose variations to improve the description and retrieval performance of the Shape Context and Generalized Shape Context descriptors. And in chapter 5 we describe a new shape descriptor that better suits for dealing with shapes containing high levels of visual complexity.

# 4 Matching of Maya Syllables

Among the different shape representations that exist in the literature, we decided to explore the use of Shape Context [Belongie et al., 2002] to describe Maya hieroglyphs. This decision is based on the good results that the method has achieved in other datasets. Another advantage of using the Shape Context is that it allows to perform matching of shapes based on local descriptors. This approach is useful when the amount of data is limited, such that any attempt to compute statistics on global descriptors for that data will result in models of poor reliability. Such was the case of the first dataset we had available at the beginning of our research, which consisted of about only 300 instances of Maya hieroglyphs.

This chapter is organized as follows. Section 4.1 describes the Shape Context algorithm, the approach to compute the point-to-point matching between shapes, and the dissimilarity score, as well as several modification we proposed to those methods and that resulted in improved results. Section 4.3 gives details about the dataset of Maya hieroglyphs used in this initial experiments. Section 4.4 explains the experimental protocol. Section 4.5 presents our results, which show that our modifications to the point-to-point matching strategy and to the computation of dissimilarity scores provide improved retrieval results. Finally, section 4.6 presents our conclusions. This work has been reported in [Roman-Rangel et al., 2009, 2011b].

## 4.1 Shape Context

This section describes how to compute the Shape Context descriptor based on the work of [Belongie et al., 2000, 2002]. We explain how to compute descriptors for a given shape, how to measure dissimilarity between two shapes, and how to find point-to-point matches between them. We also present modified approaches for dissimilarity estimation and for point-to-point matching of shapes that provide better retrieval results.

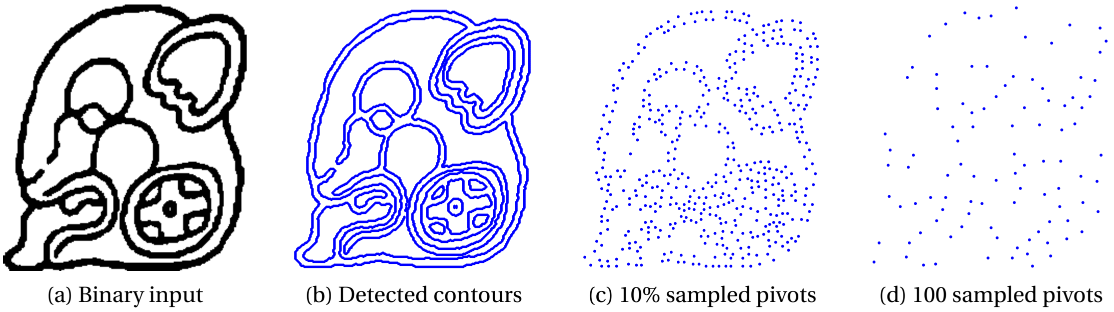


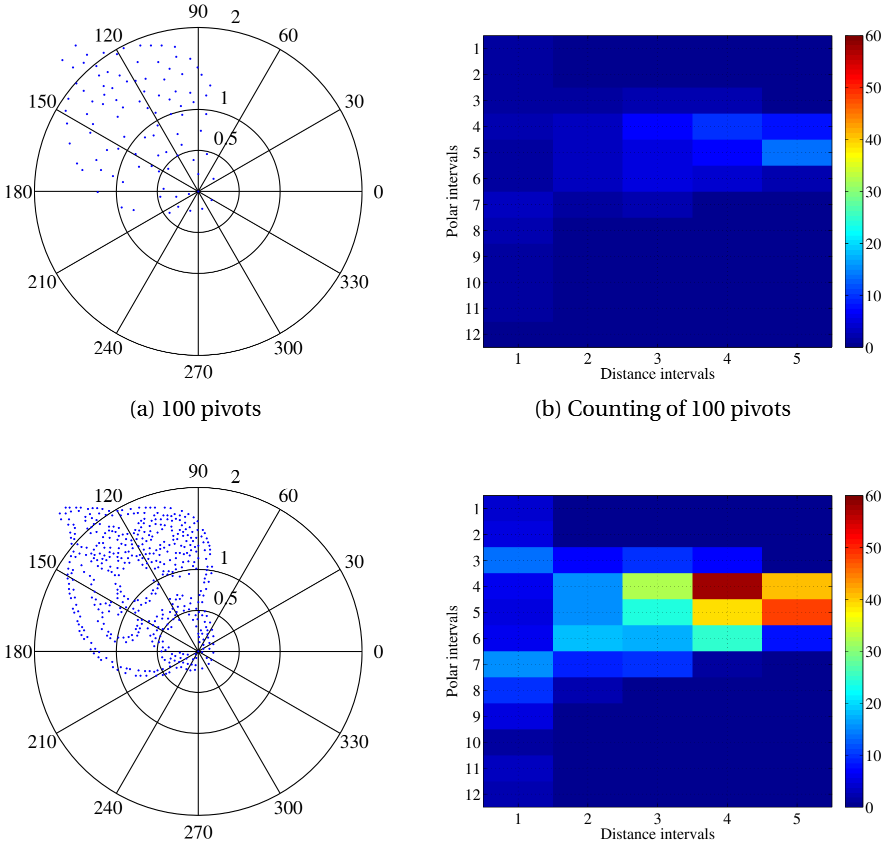
Figure 4.1: (a) Binary input of logograph B'AHLAM meaning "jaguar'; (b) contour detection; (c) 494 pivots at 10% sampling rate; (d) 100 sampled pivots, where a pivots denotes a contour point at which a descriptor is actually computed.

### 4.1.1 Descriptor

Some preprocessing is needed to set the input image into a suitable format for the Shape Context (SC) computation. The SC uses as input a binary image exhibiting the shape of interest (Figure 4.1a). Given that input, the first step consists in finding the shape contours as a set $M$ of $m$ 2-D points, as shown in Figure 4.1b. There are multiple approaches in the literature for edge detection, like Sobel [Elder and Zucker, 1998, Szeliski, 2010] or Canny [Canny, 1986] edge detectors. In our work we have used the Sobel edge detector as it produces reliable enough contours from binary images.

After the contours have been detected, a sampling process is applied to select a subset $N$ of $n$ points from the total set $M$ that represents the contours. In practice, descriptors only will be computed for this subset, as we aim to avoid a heavy computation load. This subset is obtained through a random sampling process that uniformly chooses $n \leq m$ points over the contours. In our work we have named these sampled points as "pivot points" or simply *pivots*. It is intuitive that the larger the number of pivots, the better the representation of the contours and the more accurate the description of the shape. In [Belongie et al., 2000, 2002], $n = 100$ pivots were used for the description of each shape. However, we noticed that a fixed number is not the best choice when shapes vary in complexity and amount of internal details. Therefore, we propose to consider $n$ as a proportion of the original number of points $m$. In practice 10-15% the original number of points used as pivots is enough for a good representation of shapes [Roman-Rangel et al., 2009, 2011b]. The subsets shown in Figure 4.1c and Figure 4.1d correspond respectively to the results of sampling at a rate of 10% the number of contour points and at a fixed number of 100 pivots. Note that certain areas of the contours that are missing when using only 100 pivots get covered when the sampling rate is fixed to 10%, which in this example resulted in 494 pivots.

Finally, a local-to-global shape descriptor $h_i$ is computed for each of the $n$ pivots $p_i$ based on the relative position of the $n-1$ remaining pivots[Belongie et al., 2002]. More precisely, for each pivot $p_i$ in the image $P$, its Shape Context descriptor is the histogram $h_i$ computed



(a) 100 pivots

(b) Counting of 100 pivots

(c) 494 (10%) pivots

(d) Counting of 494 pivots

Figure 4.2: (a) Log-polar formulation of Shape Context with 100 points; (b) counting of point per region with 100 points; (c) log-polar formulation with 494 points; (d) counting of points per region with 494 points.

as the number of the $n-1$ remaining pivots placed on each of the regions $r$ of a log-polar grid, as shown in Figures 4.2a and 4.2c. This log-polar grid is constructed such that the area surrounding the pivot of interest is divided in 5 concentric intervals (*rings*) that are logarithmically spaced, and that span up to twice the *average pairwise distance* computed among all the $n$ pivots; and over 12 angles intervals (*slices*) that cover a complete perimeter of $2\pi$ around the pivot $p_i$. Therefore, the resulting histogram has 60 bins which are uniform in log-polar space. As an example, Figure 4.2a and Figure 4.2c show the set of pivots sampled at different rates for the shape of the jaguar of Figure 4.1a. These pivots are placed over the log-polar grid using the same pivot as reference for both cases, i.e., the pivot in the center of the grid is the pivot for which we want to compute the descriptor. Note that, for visualization purposes, the two closest *rings* are not shown.

As seen in Figure 4.2a and Figure 4.2c, some of the $n-1$ pivots are placed beyond the maximum distance scope. When that happens, those pivots are not considered for the description of the pivot of interest. Figure 4.2b and Figure 4.2d show two matrices that illustrate the respective counting of pivots in each of the 60 regions of the log-polar grid. The Shape Context descriptor is simply the result of putting such a matrix into a vector form.

Mathematically, the Shape Context descriptor $h_i^P$ for the pivot $p_i$ of a shape $P$ can be formulated as,

$$h_i^P(r) = \left|P_i^r\right|, \tag{4.1}$$

where $r$ is the index of the 60 log-polar regions, $|\cdot|$ is the cardinality operator, and $P_i^r$ denotes the subset of pivots falling within the $r$-th region $R_r$ with respect to pivot $p_i$;

$$P_i^r = \{p_j \in N\colon\ p_j \neq p_i, \left(p_j - p_i\right) \in R_r\}, \tag{4.2}$$

where $p_j - p_i$ means vector difference.

### 4.1.2 Shape Context descriptor distance

The matching cost $c\left(p_i, p_j\right)$ between two pivots $p_i$ and $p_j$ that are described by their shape contexts $h_i$ and $h_j$ can be computed by the linear combination of their shape contexts distance $c^{sc}$ with their difference of orientations $c^{tan}$ [Belongie et al., 2002],

$$c\left(p_i, p_j\right) = \left(1 - \beta\right) c^{sc} + \beta c^{tan}, \tag{4.3}$$

where the shape contexts distance $c^{sc}$ is computed by the use of the $\chi^2$ test statistic,

$$\begin{aligned} c^{sc}\left(p_i, p_j\right) &= \chi^2\left(h_i, h_j\right) \\ &= \frac{1}{2}\sum_{r=1}^{R} \frac{\left[h_i(r) - h_j(r)\right]^2}{h_i(r) + h_j(r)}, \end{aligned} \tag{4.4}$$

and the difference of orientations $c^{tan}$ is defined as,

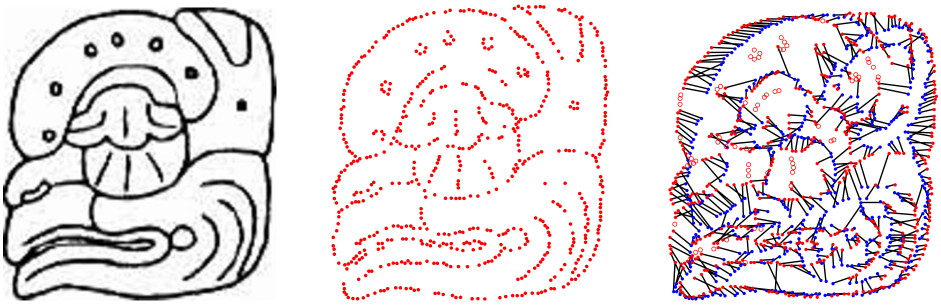
$$c^{tan}\left(p_i, p_j\right) = 0.5\left(1 - \cos\left(\theta_i - \theta_j\right)\right), \tag{4.5}$$

where $\theta_k$ denotes the local orientation of pivot $p_k$ expressed in radians.

Therefore, it is possible to construct a dissimilarity matrix $C_{ij}^{(P,Q)}$ defined between two shapes $P$ and $Q$ as,

$$C_{ij}^{(P,Q)} \equiv c\left(p_i, q_j\right), \tag{4.6}$$

where $p_i$ corresponds to the $i$-th pivot in shape $P$ and $q_j$ corresponds to the $j$-th pivot in $Q$.



(a) Binary input (b) Sampled pivots (c) Point-to-point matching

Figure 4.3: (a) binary input of syllable *ki*; (b) pivots randomly selected at 10% sampling rate; (c) point-to-point matching between shapes *ki* in (b) and B'AHLAM in Figure 4.1c. Real matches are connected by a black line while dummy matches are shown as disconnected circles.

### 4.1.3 Point-to-point matching and shape dissimilarity score

Assuming that two given shapes $P$ and $Q$ have the same number of pivots, it is possible to define a distance measure between them by finding the optimal permutation $\Pi$ of the vector of points $P$ that that minimizes

$$H(\Pi) = \sum_i C^{(P,Q)}_{\pi_i,i}, \tag{4.7}$$

where $\pi_i$ is the $i$-th element in vector $\Pi$, such that $C^{(P,Q)}_{\pi_i,i} = c\left(p_{\pi_i}, q_j\right)$, and $H(\Pi)$ defines the sum of the diagonal of the cost matrix (Equation (4.6)) after been permuted according to $\Pi$. This assignment problem can be solved by the use of the Hungarian method [Kuhn, 1955].

A constraint to this minimization problem is that the number of pivots must be the same for the two shapes. If this restriction is not satisfied, one can make use of *dummy handlers*, as done in [Belongie et al., 2002] to deal with outliers. Assuming that $|P| < |Q|$, we generate dummy pivots for $P$ until we have the same cardinality in both sets $P$ and $Q$, i.e., pivots with neither location nor description, but only entries in the cost matrix $C^{(P,Q)}_{ij}$ defined in Equation (4.6), and we fill the corresponding entries in the cost matrix with the value defined by a dummy cost $\epsilon_d$. In this way, all the pivots in $Q$ with the worst matching cost will be forced to match with dummy pivots in $P$. Figure 4.3 shows a point-to-point matching example of two shapes using dummy pivots.

In [Belongie et al., 2002], a dissimilarity score is computed between two shapes $P$ and $Q$ by,

$$d^{(P,Q)}_B = \max\left(\frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P} C, \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q} C\right), \tag{4.8}$$

where $C$ is the cost matrix after the permutation, as defined in Equation (4.7). Thus, Equation (4.8) computes the dissimilarity score as the maximum of the accumulation of the minimum point-to-point matching costs.

## 4.2 Our contributions to the Shape Context

By visual inspection of several retrieved lists of glyphs, we noticed that the Shape Context might not have enough discriminative power in some cases, for instance, when two similar glyphs have pivots whose SC descriptions encode visual structures that differ only in a few of the log-polar regions (small local changes in the shapes), resulting in a poor retrieval performance since Equation (4.8) only takes into account the minimum point-to-point matching costs. It would be more adequate to take into account the *real matching costs* obtained after the permutation of the point-to-point correspondences has been found. To alleviate this issue we propose a similarity score based on the average of the point-to-point matching costs. In pratice, this can be achieved by computing the average of the diagonal of the matrix in Equation (4.7), as suggested in [Roman-Rangel et al., 2009].

Another drawback of the SC approach is that all the pivots in the smallest set (e.g., $P$) are forced to have a real matches in the largest set $Q$. To allow outliers in $P$, we introduce dummy handlers in both sets $P$ and $Q$ by increasing the dimension of the cost matrix $C$ up to $\lceil \hat{m}(1+\rho) \rceil$, where $\hat{m} = \max(|P|, |Q|)$, and the dummy rate $\rho$ is the fraction of pivots in the largest set that are allowed to have no match. Finally, any new entry in $C$ is also set to $\epsilon_d$.

Mathematically, the combination of this augmented matrix and the computation of the similarity score as the average real point-to-point matching costs can be expressed as

$$\begin{aligned} d_{sc}^{(P,Q)} &= \frac{1}{\lceil \hat{m}(1+\rho) \rceil} \sum_{i=1}^{\lceil \hat{m}(1+\rho) \rceil} C_{\pi_i, i} \\ &= \frac{M(\Pi)}{\lceil \hat{m}(1+\rho) \rceil}. \end{aligned} \tag{4.9}$$

This function discriminates better between glyphs that have local visually variations, as it includes all the dummy matches, thus making possible to take into account their respective costs, which might have a significant contribution to the final score depending on the number of points in the two shapes to be compared.

## 4.3 Data

We used the whole Appendix 1 [Macri and Looper, 2003], which corresponds to a collection of 297 syllabic Maya glyphs. "The New Catalog of Maya Hieroglyphs, Volume 1: The Classic Period Inscriptions" [Macri and Looper, 2003] is a catalog recently compiled that represents the most comprehensive guide to the Maya symbols of the Classic Maya writing system known.

All the glyphs in the collection were saved as binary images. They were resized to fit 256 pixels in their largest axis keeping the proportion with respect to the smaller axis. From the collection, we randomly selected 22 glyphs and used them as queries. Then, we labeled as relevant instances all the glyphs that we considered visually similar to each query. Note that

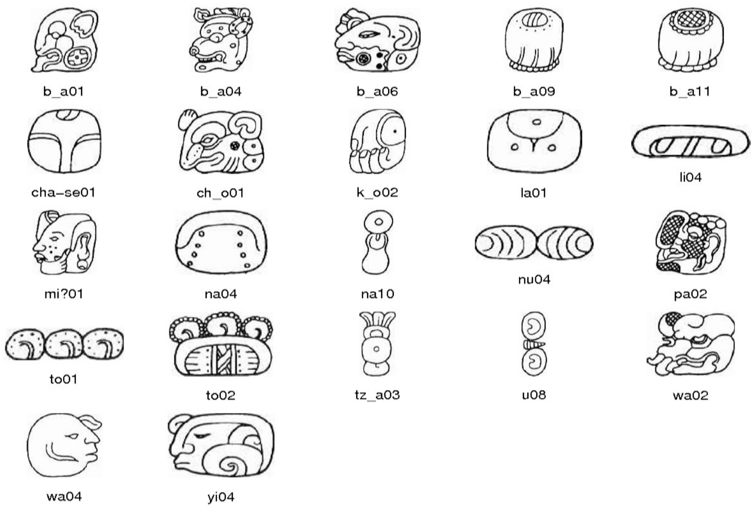


Figure 4.4: Set of 22 query syllabic Maya glyphs. The glyphs are labeled with their corresponding sounds plus a two digits consecutive number that together generate a unique identifier for each instance.

the set of queries is diverse, i.e., queries might be considerably different among each other. Figure 4.4 shows the 22 queries.

## 4.4 Experimental protocol

Since the glyphs in our dataset have different degree of details and complexity, we did an analysis to assess how well their contours are represented at different sampling rates. We tried representing the shapes with 2%, 3%, 5%, 10% and 15% the total number of points $m$ in the original contours, and observed that with ratios less than 5% many glyphs are represented by less than 100 pivots, thus yielding rather poor representations. Conversely, while 10% and higher percentages produce robust representations, they also make the SC computation slower. Empirically, 5% is a good trade-off between accurate representation and efficiency. For the experiments, we use both 5% and 10% sampling rates with a minimum bound of 100 pivots to assure a robust representations. Note that we did not used the original sampling approach (only 100 pivots), as in practice some glyphs might have contours with less that 100 points. In such a case we used all the points as pivots, that is, we used sampling rates $n = \max(\min(100, m), 0.05m)$ and $n = \max(\min(100, m), 0.1m)$.

Using Shape Context descriptors and the similarity score of Equation (4.9), we performed a series of experiments to analyze the effects of the sampling rate $n$ and the rate of dummy handlers $\rho$. We evaluated the retrieval performance in terms of mean average precision (*mAP*) [Baeza-Yates and Ribeiro-Neto, 1999], which is defined as the mean of the average precision scores achieved for each query. In turn, the average precision (*AP*) of each query is computed

| Case | $n$ | $\rho(\%)$ | $mAP$ |
|---|---|---|---|
| $a$ | max(100, 5%) | 0 | 0.301 |
| $b$ | max(100, 5%) | 10 | 0.315 |
| $c$ | max(100, 5%) | 20 | 0.319 |
| $d$ | max(100, 10%) | 0 | 0.311 |
| $e$ | max(100, 10%) | 10 | 0.319 |
| $f$ | max(100, 10%) | 15 | **0.322** |

Table 4.1: Tested combinations with Shape Context. Values for sampling rate $n$, dummy rate $\rho$, and mean Average Precision *mAP* computed over the 22 queries. The best result obtained is shown in bold.

as the average of the precision achieved by the retrieved instances that are relevant to the query, where these retrieved instances are sorted in terms of a similarity score.

## 4.5 Results

Table 4.1 shows the results for different parameter values. In all cases, the dummy assignment cost was set to $\epsilon_d = 0.25$.

It is clear how a more populated representation of the contours results in a more accurate description. Combinations *a* and *d* in Table 4.1 correspond to the original SC, with dummy assignments only for the smallest set of pivots (glyph with poorer representation) but not in the largest, from then on we can see that increasing the dummy rate in both shapes increases the precision, and the more we add the higher it is, e.g., *mAP*(a) < *mAP*(b) < *mAP*(c); likewise, *mAP*(d) < *mAP*(e) < *mAP*(f). This is due that SC now relies on a higher proportion of correct matches, assigning a dummy matching cost to the unreliable, often wrong, matches. Note that the sampling rate is higher in case *d* than in case *c*, however the precision in *c* is higher as it makes use of dummy handlers. Case *f* provides the best *mAP* at 0.322, which is more than 10 times higher than random ranking which would give a *mAP* equals to 0.030. The relative low *mAP* is due to the complex nature of the glyphs with a lot of internal contours that create confusion.

On the other hand, we noticed that increasing the value $\epsilon_d$ does not improve the *mAP* [Roman-Rangel et al., 2009], and that the use of the original similarity index $H(\Pi)$ of Equation (4.7) [Belongie et al., 2002] results in lower performance [Roman-Rangel et al., 2011b]. The relative low *mAP* in the results of Table 4.1 is due to two reasons: a) the complex nature of the glyphs which usually have a lot of internal contours that create confusion, and b) we have arbitrary assigned the relevance labels base on our perception of visual similarity.

**Individual query analysis.** In Figure 4.5 we show the standard precision-recall curves for the three queries having the highest retrieval performance (for the best combination of parameters

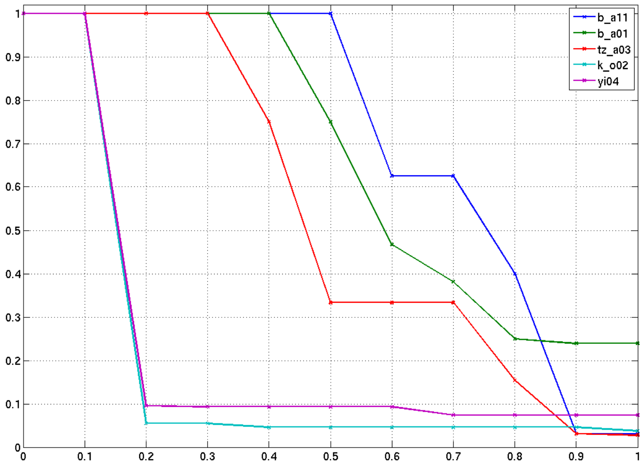


Figure 4.5: Performance details for the best combination of parameters, case *f* in Table 4.1 standard precision-recall curves for the three queries having the highest retrieval performance, and also for the two queries with the worst results. Visual results of queries and retrieved instances are shown in Figure 4.6.

*l* in Table 4.1), and Figure 4.6 displays the corresponding query-glyphs and the 5 most similar instances retrieved. As we can see, the curves of queries *b'a11* and *b'a01* remain with 100% of precision until 40% and 50% of recall. This is because the glyphs in the collection that are relevant exhibit very similar visual features, and therefore they easily get retrieved at early positions.

The third curve in Figure 4.5 corresponds to query *tz'a03*, and is worth of attention: the query has a vertical shape, as well as all the most similar retrieved glyphs, as shwon in Figure 4.6. We found the same kind of behavior for queries *na10* and *u08*, and also for horizontal shapes like those in queries *li04* and *nu04* (see Figure 4.4 for reference about the queries).

The curve for query *k'o02* in Figure 4.5, which represents a hand, performs poorly as due to the fact that we chose its relevant set under semantic criteria rather than visual similarity. More precisely, we included in the set all the hands found in the collection even those with different shapes like extended palms, fist, fist with thumb up or down, *etc.* The last curve, query *yi04*, gives poor results since it is the unique human face in the collection with its overall shape and pointing to the left, and in this case we also included faces with several shapes as relevant glyphs. However, retrieved glyphs for these two examples have similar overall shapes as seen in Figure 4.6.

In general, we noticed that the Shape Context strongly takes into account the overall shape, even when using dummy handlers that increase the rate of correct point-to-point matches of the inner details. An important restriction for this method is that the higher the dummy rate, the slower is to find the permutation Π, reason for which we stopped the exploration at rates of 20%. In chapter 5, we introduce more efficient approaches that allow for fast retrieval

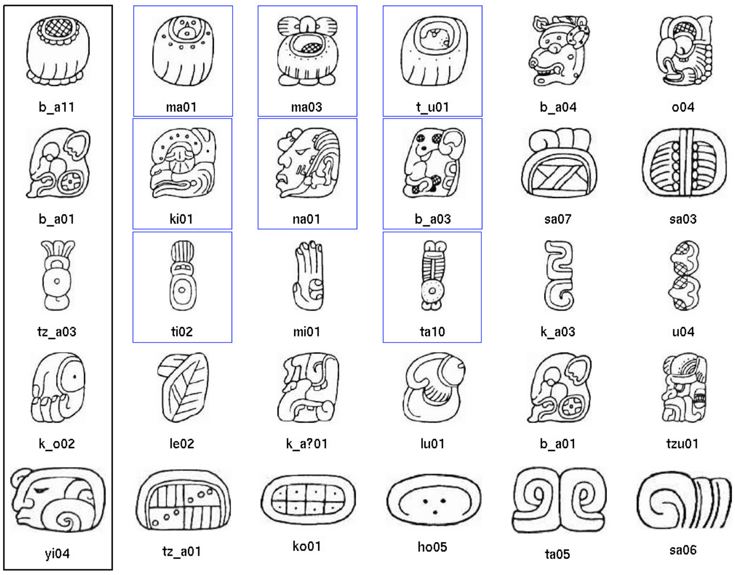


Figure 4.6: Example of retrieved glyphs. First column shows the three queries with highest *AP* and the two with lowest one (in descending order). On each row we show the top 5 retrieved glyphs, with the relevant ones been enclosed in a blue rectangle. This ranking is obtained with the best combination (case *f*) detailed in Table 4.1.

experiments.

## 4.6 Conclusions

In this chapter, as the initial step of our research, we assessed the retrieval performance of the Shape Context descriptor in an initial dataset of syllabic Maya glyphs, obtaining encouraging results. We identified sources from where we could gather data for potential use in future experiments [Macri and Looper, 2003]. We also scanned and formatted an initial dataset that we used to evaluate the performance of the Shape Context descriptor base on a set of retrieval experiments. During experiments, we found a more adaptive method to perform the sampling of the pivot-points that the Shape Context uses for description. This method allow for a better representation of the general shape structure while keeping the efficiency in terms of computation loads.

As part of our study, we improved the cost function and the dissimilarity index estimation that are traditionally used to compare sets of shape descriptors. We did that by using dummy handlers and modifying the point-to-point matching approach, which resulted in more discriminative comparison of shapes with large amount of inner details. Overall, our proposed modifications to the shape description algorithm and the comparison methodology allowed us to improve the retrieval performance of the Shape Context descriptor. In chapter 5 we will present a new shape descriptor that further improves the image retrieval performance.

# 5 Histogram-of-Orientations Shape Context (HOOSC)

The retrieval performance obtained with the original Shape Context (SC) descriptor in the experiments of chapter 4 can be improved. In this chapter, we present a new shape descriptor named *Histogram-of-Orientations Shape Context* (HOOSC). It builds on top of the Shape Context descriptor and incorporates extra information that helps generate a more robust description of the shapes. More specifically, it incorporates a distribution of local orientations in each of the regions of the polar grid of the Shape Context, which results in better retrieval results.

As mentioned in chapter 4, the point-to-point matching used by the SC can be time consuming. Therefore, in this chapter we rely on more efficient methods that use quantized descriptors, and that allow for fast comparison and retrieval. In this context, experiments with the Generalized Shape Context (GSC) [Mori et al., 2005], which is a quantized version of SC, showed to be much faster, although not as good in terms of retrieval precision. We compare the retrieval precision of the proposed HOOSC descriptor versus that of the GSC.

We conducted our experiments on a new dataset that was collected by manual identification, annotation, and segmentation of syllabic hieroglyphs that appear in Maya inscriptions of the Mexican territory. The majority of these instances have been generated by a research partner at the National Anthropology and History Institute of Mexico (INAH) through the AJIMAYA project (see chapter 2.3). The dataset was later enlarged with a few instances taken from other sources.

This chapter is organized as follows. Section 5.1 describes the Generalized Shape Context, a quantized version of SC that incorporates local orientation information. Section 5.2 describes the new HOOSC descriptor. Section 5.3 explains the quantized approach used for fast retrieval experiments. Section 5.4 introduces two new datasets used to test the shape descriptors. Section 5.5 describes the experimental protocol. Section 5.6 discusses the results obtained with our experiments. And section 5.7 concludes the chapter. The content of this chapter has been reported in [Roman-Rangel et al., 2011a,b, 2012, Gatica-Perez et al., 2011].

## 5.1 Generalized Shape Context (GSC)

The use of the contour orientation as part of the cost function between two descriptors, only at the reference pivot according to Equation (4.3), appears to be rather limited [Mori et al., 2005, Roman-Rangel et al., 2009]. In contrast, replacing the simple counting of pivots in each log-polar region $r$, by the sum of the unitary gradient vectors $\theta_j$ of all pivots $p_j$ falling within that region, provides with a richer description of shapes. This leads to the definition of the *Generalized Shape Context* (GSC) descriptor [Mori et al., 2005] $gsc_i^P$ of pivot $p_i$, in which each regions $r$ is described by an estimation of the dominant orientation $gsc_i^P(r)$:

$$gsc_i^P(r) = \sum_{p_j \in P_i^r} \theta_j, \tag{5.1}$$

where $P_i^r$ is the set of pivots falling into the log-polar region $r$ relative to the pivot $p_i$ as defined in Equation (4.2), and $\theta_j$ denotes the local orientation of pivot $p_j$ expressed in radians and normalized to unit length. For the purpose of comparing two descriptors, the dominant orientation in each region is represented as a 2-D vector $v(r)_i = \left(h(r)_i^x, h(r)_i^y\right)$ which contains its components in $x$ and $y$. As a result, using the same log-polar grid of the SC with 60 regions, the resulting GSC is a 120-D vector.

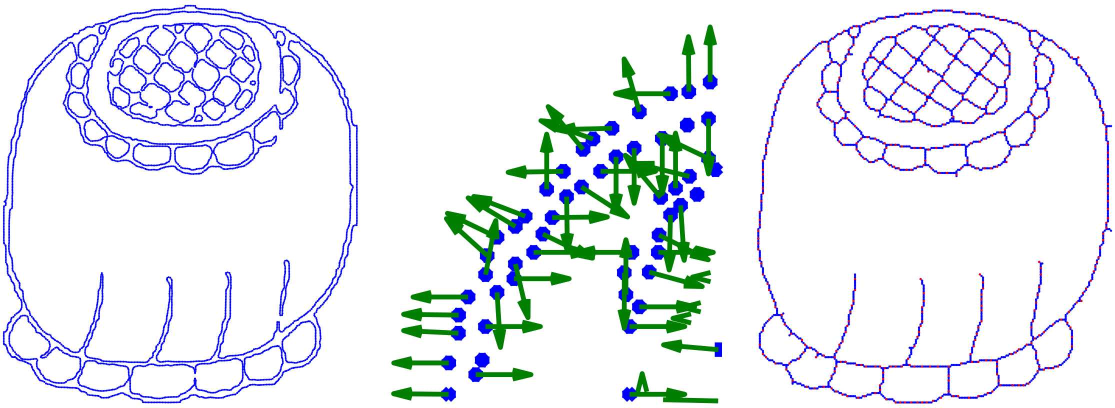
The GSC is used under a quantization approach that provides shape representations for efficient retrieval experiments [Mori et al., 2005]. However, this speedup comes at the cost of loss of spatial information, thus resulting in lower retrieval precision. In practice, GSC has been used for a fast pruning of large dataset, retrieving a set of candidate relevant instances that are then compared with the original Shape Context for accurate ranking. In section 5.2 we present a new descriptor that further exploits the local orientation information to construct robust shape representations.

## 5.2 HOOSC descriptor

This section explains the details of the *Histogram-of-Orientations Shape Context* (HOOSC), which was first introduced in [Roman-Rangel et al., 2011b], and later improved in [Roman-Rangel et al., 2011a] and [Roman-Rangel et al., 2012]. Similar to the SC and GSC descriptors, the HOOSC produces a set of semi-local descriptors for a set of pivot points that represent a given shape. Nevertheless, there are six improvements that make the HOOSC more robust than SC and GSC. Each of these improvements are described in the subsections.

### 5.2.1 Sampling of pivots from thinned shapes

The computation of shape descriptors based on pivots sampled from the contours of a raw shape works well for silhouettes and shapes whose internal details are not of crucial importance. In general terms, this approach performs well when accurate contours can be easily extracted. However, this is not the case for Maya hieroglyphs, which are rich in internal details,



(a) Detected contours

(b) Opposite gradients

(c) Thinned shape

Figure 5.1: (a) Detected contours for syllable *b'a*; (b) details of the upper left corner of (a) showing opposite gradients at each side of the contours; (c) thinned contours with sampled pivots highlighted in red color.

and that often present lines of different degrees of thickness, both along their contours as well as in their internal details.

Due to these visual variations in the Maya glyphs, the contour detectors some times generate single or "double" contours as in the example shown in Figure 5.1a, which might result in noisy descriptors and in an increase of the intra-class variability, i.e., unwanted variations in the resulting descriptors for shapes of the same visual class. This fact can lead to a potential difficulty under approaches that use local orientations, since points at opposite sides of a thick line (parallel contours) might result in opposite gradients, and therefore in opposite local orientation, whose addition cancels each other, see Figure 5.1b.

We propose the use of thinning algorithms [Lam et al., 1992] to preprocess the binary shapes and estimate thinned versions of their contours, which helps avoid potential double contours, see example shown in Figure 5.1c. The resulting thinned shapes can be seen as approximations to the underlying structure of the shape class.

### 5.2.2 Description of pivots with respect to points

Similar to the Shape Context, a sampling process helps obtain a subset of $n$ pivots from the original set of $m$ points that, in this case, represent the thinned version of the shape. However, instead of computing the shape descriptor $h_i$ for the pivot $p_i$ as a simple count of only the $n-1$ remaining pivots, we propose to compute it as a *function* of *all* the points in the original set $M$. For example, in Figure 5.1c, each pivot shown in red color will be described as a function of all the blue and red points. The resulting descriptors will be more accurate, yet will remain computationally efficient.

### 5.2.3 Histogram of local orientation of points

The use of local orientation requires the modulo $\pi$ normalization of the direction vectors. Thus we normalize them to the interval $[0,\pi)$. As a consequence of the above, the use of Equation (5.1) might provide inaccurate descriptors as modulo $\pi$ orientation vectors are difficult to add. In addition, the use of thinned shapes sometimes results in morphological variations for instances of the same class, thus generating slight variations on their local orientations that could add noise to the estimation of a dominant orientation. Furthermore, in many cases the notion of dominant orientation is not fine enough, as spatial bins might contain sets of lines with different orientations. Thus, following a successful trend in computer vision [Lowe, 2004, Dalal and Triggs, 2005], we propose a more robust approach to characterize each region, which consists of a 8-bin histogram of local orientations covering the interval $[0-\pi)$ inside each region, where the density of each bin is approximated by a truncated Gaussian kernel.

More precisely, the density of the $b$-th bin, in the portion $h_r(b)$ of the HOOSC descriptor $h_r$, for the region $r$ is estimated as,

$$h_r(b) = \sum_{p_j \in r} H_{\theta_j}(b), \tag{5.2}$$

where $\theta_j$ is the local orientation of the point $p_j$ (we need the local orientation of all the points as we describe the pivots as a function of the points), such that the summation is computed over all the points $p_j$ localized within the region $r$ (Equation (4.2)). The density function $H_o(b)$ is computed as

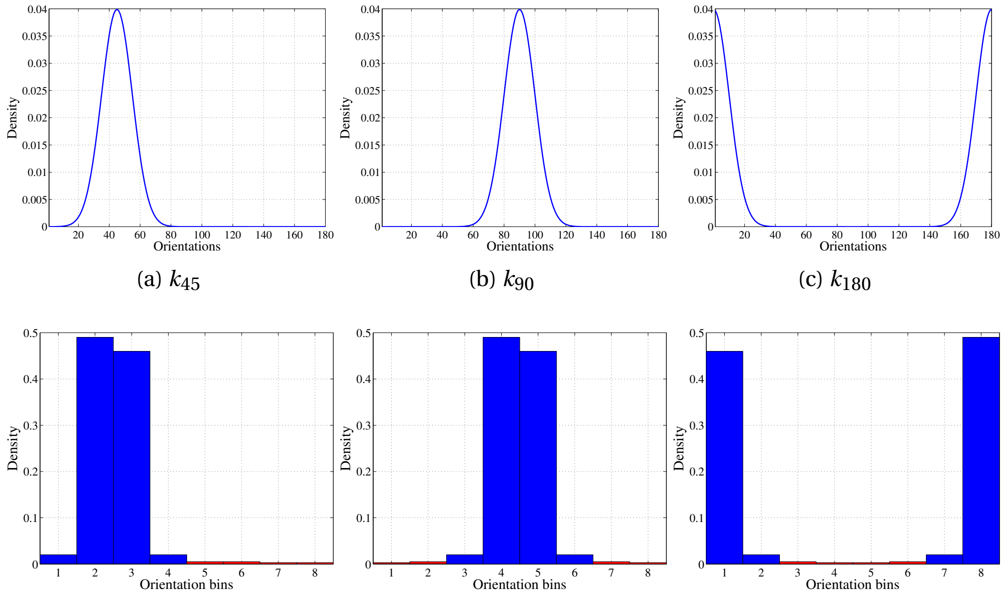
$$H_o(b) = \sum_{\theta \in b} k_o(\theta), \ \forall b = 1, \ldots, 8, \tag{5.3}$$

where $\theta \in b$ denotes all the orientation values within the $b$-th bin of the kernel $k$ used to approximate the densities. In turn, the kernel $k$ is defined as

$$k_o(\theta) = \mathcal{N}\left(\theta; \theta_o, \sigma^2\right), \tag{5.4}$$

where $\mathcal{N}\left(\theta; \mu, \sigma^2\right)$ denotes the value of a Gaussian having mean $\mu$ and variance $\sigma^2$. A value of $\sigma = 10$ has shown to work well, avoiding hard binning effects and dealing with imprecision in orientation estimation [Roman-Rangel et al., 2011a]. Figure 5.2 shows the Gaussian kernels $k_{45}$, $k_{90}$, and $k_{180}$ for the respective angles of 45°, 90°, and 180°, and their corresponding density functions $H_{45}$, $H_{90}$, and $H_{180}$.

The Gaussian density functions $H_o$ have tails whose densities are very close to zero and that can be considered as “noise". To eliminate that noise while keeping the advantage of such an efficient method against hard binning effects and imprecision in orientation estimation, we propose to use a truncated Gaussian assumption, i.e., we set to zero the 4 bins in each Gaussian density function corresponding to its smallest values. Therefore, the histogram of local orientations in each region is computed as the summation of only the 4



(a) $k_{45}$ (b) $k_{90}$ (c) $k_{180}$

(d) $H_{45}$ (e) $H_{90}$ (f) $H_{180}$

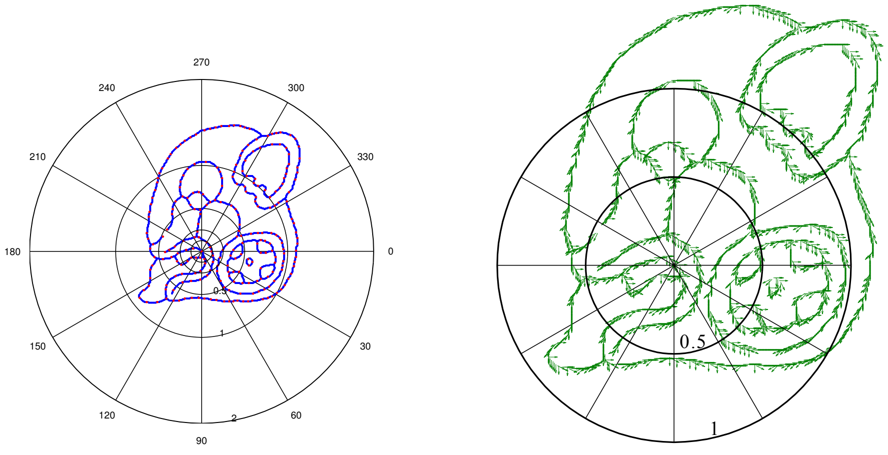
Figure 5.2: Gaussian kernels $k_{45}$, $k_{90}$, and $k_{180}$, and density functions $H_{45}$, $H_{90}$, and $H_{180}$, used to approximate the density of the local orientations for the HOOSC descriptor. Using a truncated Gaussian model, we set to zero the less informative intervals (red) in the tails of the *pdf*'s, and used only the most central bins (blue).

most probable orientation bins for each of the points within that region, i.e., the truncated Gaussians only contribute with their respective 4 most representative bins and add nothing to the less probable bins. The Gaussian densities shown in Figure 5.2 have their 4 most representative bins in blue and their “noisy tails" in red.

### 5.2.4 Bounded distance context

Using 8-bins histograms of orientations in each of the 60 regions of the log-polar grid, will result in 480-D vectors. This is the case of the original HOOSC descriptor reported in [Roman-Rangel et al., 2011b]. However, we noticed that the most internal regions of the log-polar space usually include very few points (sometimes only the pivot to be described is in those regions). Also very often, many of the external regions are empty or only contain points that are close to the inner boundary of this distance inverval. These two facts might result in several empty or noisy sections of the 480-D HOOSC vector. Therefore, there is no need for using all the 5 distance intervals (*rings*) and covering twice the pairwise distance among points (see chapter 4.1). There are two variants of the HOOSC descriptor reported in [Roman-Rangel et al., 2011a, 2012] that use different spatial context formulations around the pivots of interest:

1. In [Roman-Rangel et al., 2011a] the HOOSC uses only a intermediate spatial scope formed by the second, third, and fourth rings. However, there might be information



(a) log-polar grid

(b) simplified polar grid

Figure 5.3: (a) Log-polar grid of the SC, the plot highlights the pivots in red color. (b) Simplified grid used with the HOOSC. It shows the local orientation of each point in (a). Note that all the points beyond the average pairwise distance are not used to build the HOOSC descriptor, as compared to two times the average distance used for the Shape Context descriptor.

regarding the most local area around the reference pivot that would be lost by using this constrained spatial context. This descriptor is a 288-D vector.

2. In [Roman-Rangel et al., 2012] the spatial split of the polar region is reformulated by using only two rings: one for the interval $[0-0.5)$ (thus, merging the first three rings of the log-polar grid into a single one), and other for the interval $[0.5-1.0]$ times the average pairwise distance of the input set. This spatial context organization retains the orientation information of the closest points. In practice, this local spatial scope contains the regions that are the most informative for a generic shape pivot. This results in a polar grid with only 24 regions $r = 1, \dots, 24$. This descriptor is a 192-D vector.

Note that in both configuration the last ring is dropped, therefore increasing the degree of locality. Figure 5.3 shows a comparison of the original log-polar grid of the SC and the simplified polar grid used with the HOOSC in [Roman-Rangel et al., 2012].

### 5.2.5 Normalization

There exist different options for normalization of the resulting HOOSC descriptor, i.e., the whole at once, independently per ring, independently per polar division (*slice*), independent per region, or combinations of them. We decided to apply a per ring normalization, as in practice it resulted in better descriptions compared to a global or a more specific normalization schemes. Note that after this normalization, the summation of the HOOSC entries will be equal to the number of rings used for description, except if one of the rings contains no points.

Note that in [Belongie et al., 2002] no normalization is mentioned for the SC, while in [Mori et al., 2005] the normalization is performed once for all the regions in the GSC.

### 5.2.6 Explicit relative position of the pivot in its own description

Both the SC and the GSC implicitly encode continuous information about the position of each pivot within the shape. For instance, few observations in the lower regions of the log-polar grid means that the pivot is located towards the bottom of the shape image.

However, their descriptive ability can be improved further if the position is explicitly incorporated within the descriptor. The HOOSC approach concatenates the coordinates $\left(x_i, y_i\right)$ of each pivot $p_i$ as two additional dimensions in the descriptor, thus representing the relative position within the bounding box encapsulating the glyph, i.e., within the interval $[0, 1]$. Despite being one-dimensional, each of these normalized coordinates provide approximately the same amount of information as each of the normalized rings.

### 5.2.7 Summary of the HOOSC descriptor

In summary the HOOSC descriptor is computed by:

1. Thinning the contours of the shape.
2. Computing local orientation of each point in the thinned shape.
3. Subsampling the pivots for description.
4. Placing each pivot in the center of a polar grid, and computing a histogram of orientations for each cell of the grid.
5. Concatenating the cell histograms into a vector form.
6. Normalizing the resulting vector independently for each of the spatial intervals of the polar grid.
7. Concatenating the relative position of the reference point to its own description.

## 5.3 Bag of Visual Words

In chapter 4.1.3, we explained a point-to-point matching approach used to compare sets of shapes descriptors. This is an instance of the assignment problem that is solved by the Hungarian method [Kuhn, 1955], whose computational complexity is of order $O\left(n^3\right)$. Thus, it results unfeasible to use as the datasets or the dimensionality of the descritors grow.

To perform efficient comparisons, we rely on a quantization approach known as bag-of-words (or bag-of-visterms, i.e., visual terms, *bov*), which is widely used in the image retrieval community as it has shown to allow the design of fast retrieval applications [Sivic and Zisserman, 2003, Mori et al., 2005, Quelhas et al., 2005, 2007] (the term *shapemes* is used in [Mori et al., 2005]). Under this approach, documents are represented as a simple counts of prototype-terms (words in the case of text documents) defining a so-called vocabulary.

This approach has been successfully generalized to different types of data such as images [Sivic and Zisserman, 2003], where documents are represented by local image descriptors or patches instead of text words. Since local image descriptors contain continuous values, the generation of a finite vocabulary requires a quantization process, in which the prototype-terms are first estimated, and then all the local descriptors are assigned to one or more of these bases [Quelhas et al., 2005]. As a result of this quantization procedure, the final representation of a set of points is a vector of fixed size (*vocabulary* size) containing the frequency of each *visual word.* Often this vector is normalized to actually represent the words distribution.

In the following, we review the *k*-means algorithm [Lloyd, 1982, Baeza-Yates and Ribeiro-Neto, 1999], which is one of the most common methods used to estimate visual vocabularies (also called dictionaries) and quantize local descriptors.

### 5.3.1 The *k*-means algorithm

For a given a set $X$ of $I$ input signals $x_i$ (e.g., local image descriptors), $k$-means estimates the column elements (bases) of the dictionary matrix $D = [d_1, d_2, \ldots, d_K]$ by looking iteratively for clusters $c_j = \{x_i | g(x_i) = j\}$, where $g(\cdot)$ denotes the cluster assignment function, such that the square of the euclidean distance of each descriptor $x_i$ to the center of its respective cluster $d_j$ (basis) is smaller than the distance to any other center $d_k$:

$$g(x_i) = j \iff \|x_i - D\omega_i^j\|_2^2 \leq \|x_i - D\omega_i^k\|_2^2, \forall k \neq j, \tag{5.5}$$

where $\|\cdot\|_2^2$ denotes the square of the $l_2$ norm, and $\omega_i^j$ is the unit weight row vector with its $j$-th entry set to one and the rest to zero, and it is associated to the signal $x_i$. In other words, the problem consists in finding the solution to,

$$\min_{D,\Omega}\{\|X - D\Omega\|_F^2\} \qquad \text{s.t.} \quad \forall i, \|\omega_i\|_0 = 1, \tag{5.6}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm, $\|\omega_i\|_0$ is the $l_0$ pseudo-norm defining the number of non-zero entries in $\omega_i$, and $\Omega$ is the matrix of weight row vectors $\omega_i$.

Note that allowing $\omega_i$ to be a normalized vector with more than one non-zero entry corresponds to a weighted fuzzy assignment to more than one cluster [Nock and Nielsen, 2006].

The $k$-means algorithm is an instance of the Expectation-Maximization (ME) algorithm [Dempster et al., 1977], in which the E-step corresponds to the estimation of the centroids of the clusters, whereas the M-step corresponds to the assignment of each training signal to one of the estimated clusters. The algorithm aims to minimize the sum of squares distances computed from each vector in the cluster to its centroid. This EM process iterates until it converges. A common option to chose the initial set of cluster centers, is to use a random subset of the input signals. Later in each iteration, the cluster centers are recomputed as the component-wise mean of all the descriptors within each cluster.

### 5.3.2 *k*-means with the $l_1$ norm

During our work, we observed that a variant of the *k*-means algorithm that uses the $L1$ distance performs better than the 'euclidean' distance. This variant computes the distance between two vectors, needed for the cluster assignment of the descriptors, by using the $l_1$ norm, and the centroid of each cluster is estimated as the component-wise median (rather that the mean) of all the points within the cluster. We use this variant for our experiments through this document, keeping the name as *k*-means for simplicity, unless otherwise stated.

### 5.3.3 Comparing Maya shapes under the *bov* approach

After the quantization of a set of points into a single vector, images can be easily compared by computing distances between the resulting *bov* vectors. To perform shape retrieval, we rank the candidate-shapes according to the $l_1$ distance that is computed for their *bov* with respect to the *bov* of a given query-shape. More precisely, the distance $d_{bov}$ between the *bov*'s for shapes $P$ and $Q$ is computed as,

$$\begin{aligned} d_{bov}(P,Q) &= \left|\left|bov^P - bov^Q\right|\right|_1 \\ &= \sum_{v=1}^{V} \left|bov^P(v) - bov^Q(v)\right|, \end{aligned} \tag{5.7}$$

where the summation is computed over $V$ visual words, i.e., the size of the vocabulary.

## 5.4 Hieroglyphic Data

To evaluate the retrieval performance of the different shape descriptors we tested them on two datasets, one of which was used to conduct a statistical analysis of the evolution of characters over time and across regions of the ancient Maya world.

### 5.4.1 Syllabic Maya dataset (SM dataset)

The first dataset is a compiled set of Maya hieroglyphs, whose gathering responds to the goal that partially motivates our research; that of fulfilling the need of the archaeological community of an efficient machine that help them deal with the visual complexity of the hieroglyphs. The instances in this dataset correspond to glyphs that appear often in stone

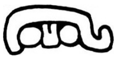
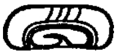
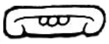
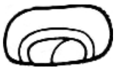
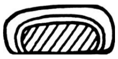
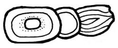
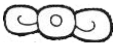
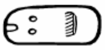
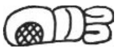
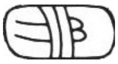
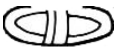
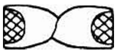
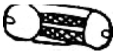
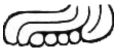
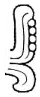
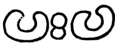
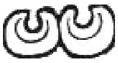
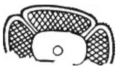
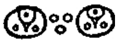
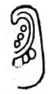
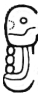
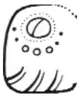
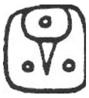
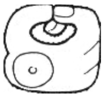
| | | | | | |
|---|---|---|---|---|---|
| T1<br><br>*/u/* | T17<br><br>*/yi/* | T23<br><br>*/na/* | T24<br><br>*/li/* | T25<br><br>*/ka/* | T59<br><br>*/ti/* |
| T61<br><br>*/yu/* | T82<br><br>*/li/* | T92<br><br>*/tu/* | T102<br><br>*/ki/* | T103<br><br>*/ta/* | T106<br><br>*/nu/* |
| T110<br><br>*/ko/* | T116<br><br>*/ni/* | T117<br><br>*/wi/* | T126<br><br>*/ya/* | T136<br><br>*/ji/* | T173<br><br>*/mi/* |
| T178<br><br>*/la/* | T181<br><br>*/ja/* | T229<br><br>*/'a/* | T501<br><br>*/b'a/* | T534<br><br>*/la/* | T671<br><br>*/chi/* |

Table 5.1: Thompson numbers, visual examples, and syllabic values (sounds) for the 24 classes of the syllabic Maya dataset.

inscriptions from 4 main subregions of the Maya area, i.e., Petén, Usumacinta, Motagua, and Yucatán. All of the instances in this dataset are syllabograms that belong to one of the 24 most common syllabic classes of the Maya writing system. The reason to use these 24 classes relies primarily on their higher frequency of occurrence over other syllabic signs, thus facilitating the manual localization and extraction of their instances by experts, and allowing us to potentially have enough material for experimentation

More specifically, this dataset is composed by roughly 900 instances extracted from inscriptions that the project AJIMAYA collected, plus another 330+ glyphs taken from [Macri and Looper, 2003, FAM, Thompson, 1962], thus reaching a total of 1270 images distributed over the 24 classes. The indispensable participation of archaeological researchers helped validate the localization and segmentation of each instance. Finally, each glyph was manually rotated to the most common orientation usually seen on its class. Table 5.1 shows one visual example for each of these 24 classes, along with their Thompson number [Thompson, 1962] which is traditionally used as identifier, and their syllabic value, i.e., their sound. Note that different to the dataset used in chapter 4, the SM dataset was compiled by experts in epigraphy. Such that its instances are not arbitrary grouped only base on our visual criterion, but rather based on visual criteria with archaeological support. In practice, the SM dataset was complemented with instances from the dataset used in chapter 4 whose classes are defined in the SM dataset.

As simple as it might sound, this compilation task required non-trivial work of expert archaeologists in Mayan iconography, who spent several months looking manually for the images in complex inscriptions. A dataset like this cannot be produced by non-trained annotators.

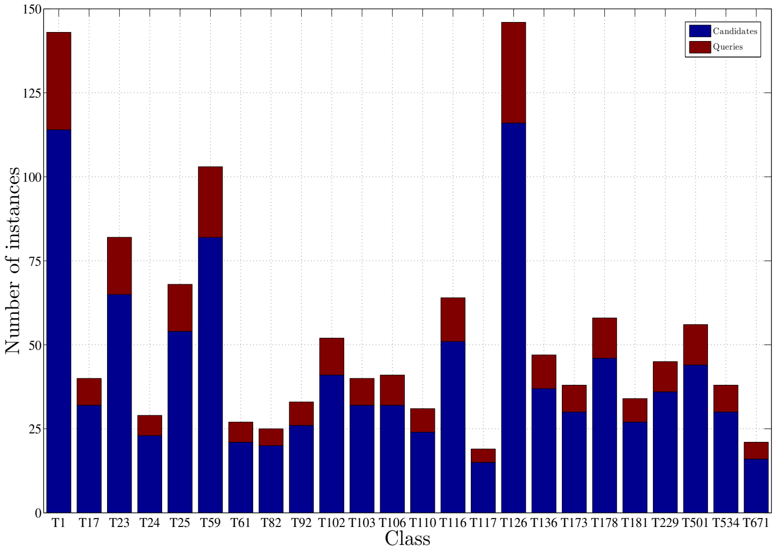


Figure 5.4: 1270 glyphs distributed over 24 classes. Number of candidates in blue and number of queries in red.

This dataset also includes 2100+ glyphs that do not belong to any of the 24 popular classes and that are grouped into a negative class. The negative instances were gathered from the same sources, taking at random as many glyphs as possible. Note that some of the glyphs in this negative class might be logographs (see section 2.2).

This dataset posits many challenges in terms of visual complexity due to the richness in internal details of its elements, their variability, the fact that some of the classes might be visually similar, and that conversely some glyphs inside each class might not be as similar as expected in visual terms. To the best of our knowledge, this is the largest dataset of Maya glyphs that has been analyzed with automatic techniques.

We divided the dataset into two subsets. Approximately 80% of the glyphs from each positive class are in the first subset (called candidates, and denoted by $G_C$), comprising 1004 instances and leaving the remaining 266 glyphs (≈20%) in the second subset (called queries, and denoted by $G_Q$). Figure 5.4 shows the number of glyphs in each class. Note that the classes are not balanced, which is a natural challenge given the structure of the Maya writing system.

### 5.4.2 INAH dataset

The second dataset consists of a subset of the syllabic Maya dataset, described in the previous subsection. It was defined by our partner from INAH who constructed tables that arrange different instances of glyphs with visual variations along temporal and regional axes.

The synchronic or time dimension is divided in 3 periods of the Maya civilization: period 1 corresponding to the Early Classic (200 - 500 AD), period 2 to the Late Classic (600 - 800 AD), and period 3 to the Terminal Classic (800 - 950 AD). The diachronic or regional dimension is divided in 4 main regions: the central region (Petén), the southern region (Motagua), the western region (Usumacinta), and the northern region (Yucatán) (see map in Figure 2.1).

These tables were built for the 8 syllabic classes shown in Figure 5.5. The reason to use only 8 tables is because these classes are fairly popular, and therefore, it resulted easier to populate the 12 entries in the tables. In practice, this was not always possible as some glyphs were never used at certain regions or times, or they have not been discovered yet. In total, this dataset contains 84 glyphs distributed over the 8 classes.

Although a majority of syllables could be arranged in such a tabular form, it was advised by our INAH partner to focus the analysis on signs with a very high rate of occurrence within the hieroglyphic corpus, thus yielding tables with as few empty entries as possible. As simple as it might sound, the selection process can often prove difficult and time consuming, as it relies almost entirely on specialized epigraphic knowledge and abilities that take many years to develop, including visual memory. It is precisely because of these difficulties that the field of Maya studies need research tools which could automatically or semi-automatically retrieve such candidates.

## 5.5 Experimental Protocol

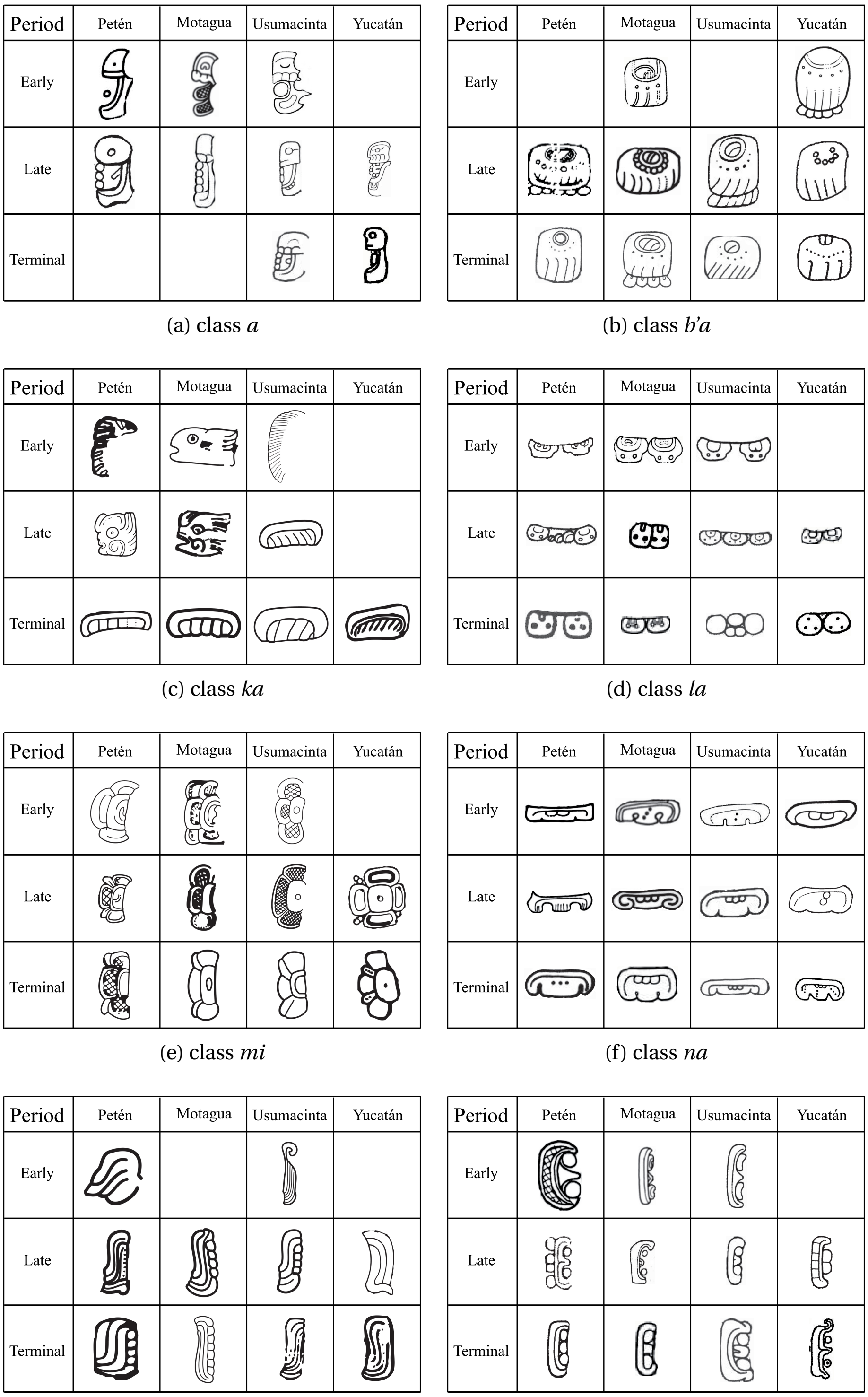
In this section we describe the experimental protocol followed to evaluate the GSC and HOOSC descriptors using a *bov* approach.

### 5.5.1 Evaluated methods

We evaluated the Generalized Shape Context (GSC) [Mori et al., 2005] and the five different variants of the HOOSC descriptor shown in Table 5.2.

The variant HOOSC0 takes points from potentially thick contours as input, describes each pivot as a function of only the other pivots, uses the log-polar grid with five rings, and does not include the relative self-position of the pivots. Starting from the definition of HOOSC0, the improvements are as follows. HOOSC1 takes as input thinned versions of the shape instead of the thick contours. HOOSC2 describes the pivots with respect to the whole set of points. HOOSC3 restructures the log-polar grid into a simplified polar grid with only two rings. Finally in HOOSC4, the relative self-position is explicitly incorporated within the description.

As a consequence of the above mentioned combinations of options, the vectors named as HOOSC3 and HOOSC4 have 192 and 194 dimensions respectively, instead of 480 as is the case of the HOOSC0, HOOSC1, and HOOSC2.



(a) class *a*

(b) class *b'a*

(c) class *ka*

(d) class *la*

(e) class *mi*

(f) class *na*

(g) class *ni*

(h) class *u*

Figure 5.5: Tables for the 8 syllabic classes of the INAH dataset. The instances correspond to three periods and four regions of the ancient Maya civilization. Missing instances either could not be found or simply would require much time to be located within the known corpus.

| HOOSC | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Contours | Thick | Thin | Thin | Thin | Thin |
| Pivots w.r.t. | Pivots | Pivots | Points | Points | Points |
| Rings | 1:5 | 1:5 | 1:5 | 1:2 | 1:2 |
| self-position | NO | NO | NO | NO | YES |
| dimensionality | 480 | 480 | 480 | 192 | 194 |

Table 5.2: The five HOOSC variants evaluated to retrieve instances of the syllabic Maya dataset. The first row indicates a version of the HOOSC descriptor. We use that number to refer to the versions as HOOSCX (e.g., HOOSC4). Improvements are highlighted in blue. Note that when using only rings 1 and 2, it implies a reformulation of the distance intervals of the polar grid (see chapter 5.2). The last row indicates the dimensionality of the resulting vector.

### 5.5.2 Evaluation on the syllabic Maya dataset (SM dataset)

With GSC and HOOSC0, we use one tenth of the number of points on the input contours, constraining it when possible to be at least 100 points, i.e., $max(min(100, m), 0.1m)$. In contrast, sampling from thinned contours usually results in less points than sampling from the double lines generated by the raw contour. To avoid very sparse sets of pivots when sampling from thinned shapes, we increased the sampling rate to $max(min(150, m), 0.2m)$, obtaining on average the same number of pivots per glyph than in the raw contour cases: 161.7 and 169.5 respectively.

From the subset of candidate-glyphs (denoted $G_C$), we randomly selected 1500 descriptors from each of the 24 positive classes, and clustered them into 2500 visual words using *k*-means. Then, we estimated the *bov* of each glyph in $G_C$ and $G_Q$. Finally, we performed retrieval experiments querying each glyph in $G_Q$ versus all the elements in $G_C$, and evaluated the retrieval precision. We reported the results in terms of mean Average Precision (*mAP*). More precisely, for each method mentioned in Table 5.2 we implemented the following protocol:

1. Compute the descriptors, (i.e., GSC or HOOSC#).
2. Estimate a visual vocabulary, using only descriptors in $G_C$ of the 24 classes. This ensures that the vocabulary does not contain information about the queries in $G_Q$.
3. Describe every glyph in $G_C$ and $G_Q$ as a *bov* distribution over the resulting vocabulary.
4. Query from $G_C$ using each glyph of $G_Q$, rank the retrieved vector, and compute the mean average precision.

To test the shape descriptors using more non-relevant glyphs, we repeated the retrieval experiment adding to $G_C$ all the glyphs of the negative class, and representing them by their *bov* computed over the visual vocabulary. This experiment is referred to as “24 + N”.

### 5.5.3 Assessing glyph variability across periods and regions

The second task we performed consists in analyzing trends of visual variability of Maya syllabograms. Such an evaluation was done both at the intra and inter-class levels with the goal of finding out whether there are visual classes showing more consistency across historical periods and regions, as well as being able to numerically characterize visual similarity among the glyphs.

We consider this evaluation important as it addresses one of the needs of scholars working on archaeological and epigraphic related studies; they look for tools to help them classify known and new hieroglyphs, as well as a better understanding of the visual relationship among instances of the same sign and with respect to other signs. Examples of these needs can be found in two seminal taxonomical and paleographical contributions conducted by [Grube, 1989] and [Lacadena, 1995], which constitute efforts to account for the amount of synchronic and diachronic variability found within the Maya script.

## 5.6 Results

In this section, we present the results obtained on retrieval experiments with the syllabic Maya dataset. We then present the synchronic and diachronic analysis for the 8 classes presented in the tables of Figure 5.5.

### 5.6.1 Retrieval of Maya syllables

The first row of results in Table 5.3 shows the mean Average Precision (*mAP*) of each method evaluated using the 24 positive classes of the Syllabic Maya dataset. The original HOOSC method [Roman-Rangel et al., 2011b] (HOOSC0) obtains a precision 12% higher than the GSC in absolute terms. Changing the input to be thinned estimations of the shapes makes the description more robust and leads to better retrieval results (HOOSC1). The *mAP* of HOOSC2 shows that taking into account all the points for the description improves the results. We also noticed that computing descriptors for the whole set of input points does not provide substantial improvements. After restructuring the distance intervals of the polar grid, and using only the closest distance scope (rings 1, 2) as in the case of the HOOSC3, the resulting descriptors are shortened while leading to a slightly higher precision. Finally, the explicit addition of the self-position (HOOSC4) allows for a *mAP* result of 0.538, for a total improvement of almost 18.8% in absolute terms with respect to the original HOOSC.

Using the GSC, we also conducted retrieval experiments only on the subset of glyphs used in Chapter 4, such that the comparison with the Shape Context (SC) method could be possible. These experiments resulted in a *mAP* of 0.279, suggesting that the SC method provides with better description and retrieval results (0.322 as seen in Table 4.1) in comparison to the GSC. Such an observation should not be surprising, as the use of the dominant orientation results

| Classes | GSC | HOOSC | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| 24 | 0.236 | 0.350 | 0.422 | 0.492 | 0.502 | **0.538** |
| 24 + N | 0.195 | 0.201 | 0.281 | 0.341 | 0.341 | **0.374** |

Table 5.3: *mAP* for the 6 methods evaluated in the SM dataset: GSC, and HOOSC0 to HOOSC4.

in the lost of certain local detail of the visual structure of the shapes. According to the proposal of the GSC [Mori et al., 2005], this method was used only for fast selection of the subset of the most similar elements in a collection, the actual ranking and evaluation performance was then conducted on this subset using the point-to-point matching of the SC.

Besides improving the *mAP*, the *bov* approach is much faster than SC. Given a query image, the quantized approach needs to compute a set of descriptors (GSC or HOOSC) and the corresponding *bov*, and then compare it against each of the *bov* of the elements already indexed in the dataset. On the other hand, the point-to-point matching approach requires to calculate the set of point descriptors, and then find the corresponding point-to-point matches. On average and using non-optimized Matlab code, for a given query the *bov* approach takes only 0.05 seconds for the whole process of description and comparison, while the point-to-point matching approach requires 227.4 seconds (3'48"). Therefore, using the *bov* approach results in a speedup factor of ≈ 4500 using a standard desk- top machine.

Since the classes might vary in terms of size and visual complexity, in Figure 5.6 we present the per-class average precision *AP* versus the standard recall for the 5 classes with the highest *AP* and the 3 with the lowest *AP* (see also Table 5.4). Although class T117 has very few instances (see Figure 5.4), it is the one with the highest average precision. This is because it contains unique features that are not shared with any other class, such as its vertical orientation and the circles in the right hand side. Similar trends occur with classes T534 (inverted face), T229 (one circle in a superior section, and some circles in a vertical arrangement on the left hand side), T59 (concentric circles and quasi-parallel lines), and T501 (circles and lines in specific internal regions).

In contrast, the curve of class T136 degrades relatively fast because its instances are often confused with class T126 (see Table 5.4). We observed a similar behavior with class T24 which is confused with classes T1, T17, and T23. In the case of class T106, the high variability among their instances, which could be split into two visual subclasses, results in a relative low precision. Despite the relative low precision for few classes, note that on average the precision is acceptable as shown in the dashed line in Figure 5.6, and in the examples of Table 5.4.

Finally, the second row of results in Table 5.3 presents the *mAP* when the 2128 elements in the negative class are incorporated within the pool $G_C$. This is a much more challenging case and the overall drop of performance for all methods reflects this fact. At the same time, note that the performance improvement over the different methods follows the same trend as the results shown in the first row of Table 5.3. Moreover, it is the same method that achieves the

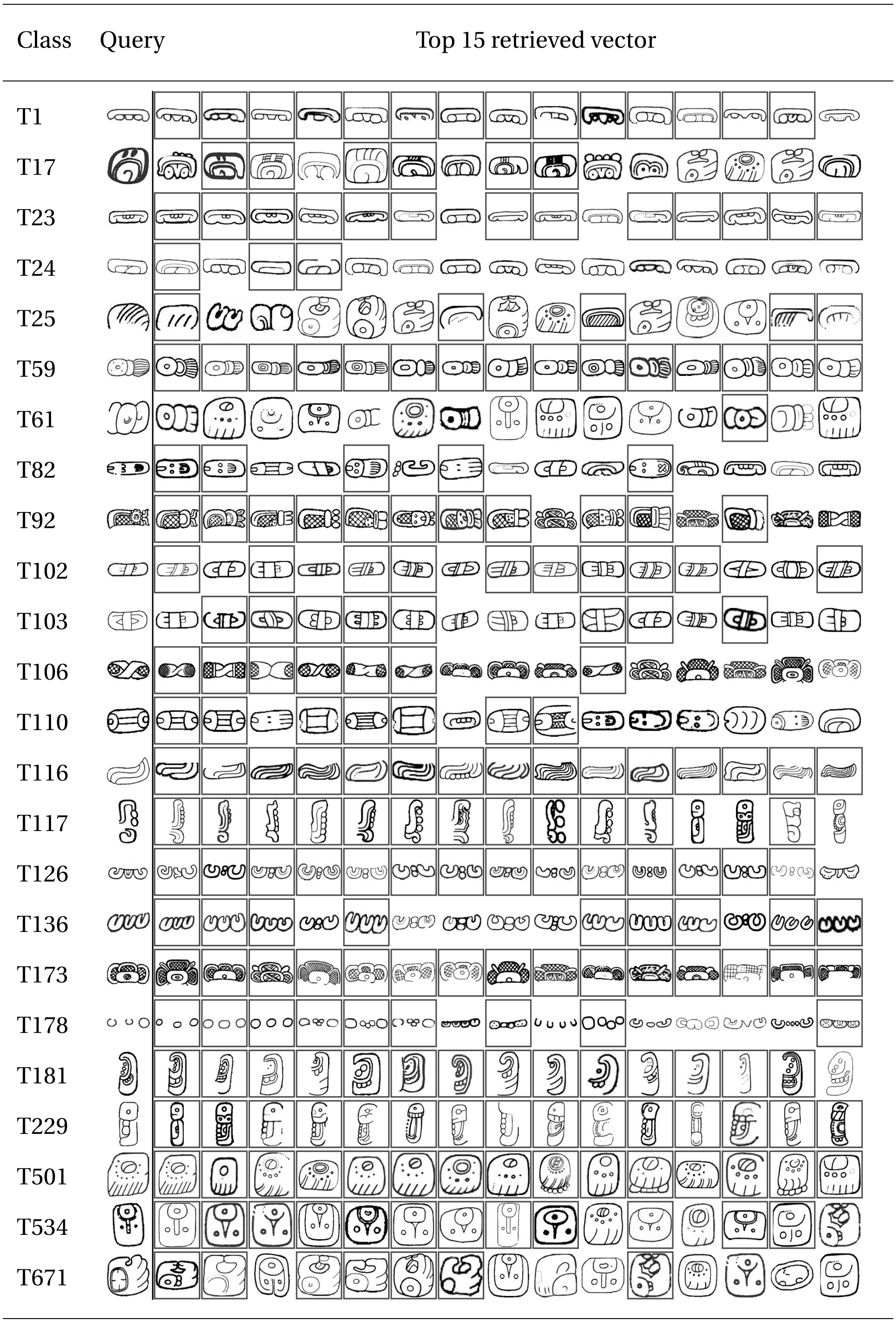

Class
Query
Top 15 retrieved vector
T1
T17
T23
T24
T25
T59
T61
T82
T92
T102
T103
T106
T110
T116
T117
T126
T136
T173
T178
T181
T229
T501
T534
T671

Table 5.4: Retrieval results on the SM dataset using HOOSC4. The first and second columns show the name of each class and one random query, followed by its Top 15 retrieved candidate-glyphs sorted in descending order of similarity from left to right. Relevant glyphs are enclosed in a gray square.

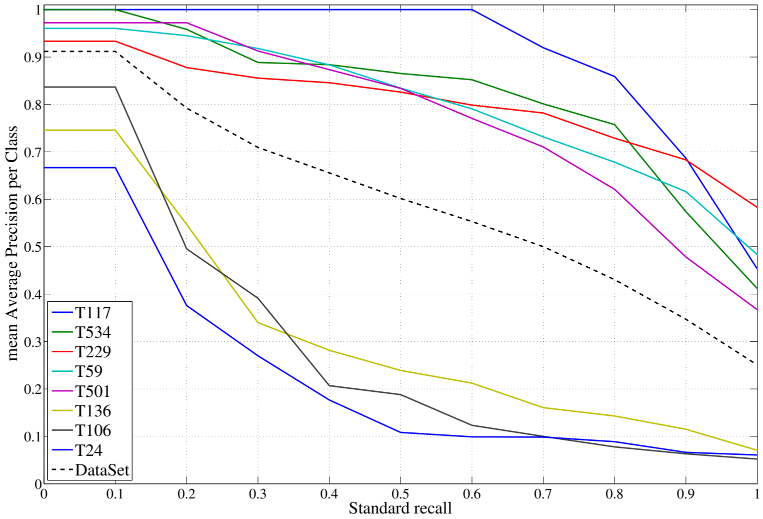


Figure 5.6: *mAP* precision vs standard recall for the whole collection (dashed line), plus the corresponding results for the 5 classes with highest average precision and the three with lowest average precision.

best retrieval results in both cases, i.e., HOOSC4, with 17.3% absolute improvement over the original HOOSC, and 17.9% over the GSC.

### 5.6.2 Intra-class and Inter-class Similarity

We now present the analysis of the visual variability of syllabic classes of the INAH dataset. This analysis was performed by computing the pairwise distance $d_{bov}$ (Equation (5.7)) between all the 84 instances, and then computing average and variance of them according to intra or inter-class criteria.

#### Intra-class analysis

The first result of our analysis is that the class *mi* (T173) (see Figure 5.5) has the lowest visual variability (average of 0.192), which is consistent with its good retrieval results (Table 5.4). Conversely, class *k'a* (T25) presents the maximum visual variability with a mean distance of 0.300. This can be explained mainly by the presence of two visual subclasses, as shown in Table 5.5. The rest of the classes have very similar distance values (≈ 0.23), meaning that most of the them have a similar degree of intra-class variability.

#### Analysis over historical periods

The second column of Table 5.5, labeled "highest", shows the list of the syllables that reached their highest intra-class variability on a given period. The last two columns of Table 5.5 show

| | Syllables | | **Variability** | |
|---|---|---|---|---|
| **Period** | **highest** | **lowest** | **average** | **std** |
| Early Classic | *a, ka, mi, ni, u* | *la* | 0.277 | 0.063 |
| Late Classic | *la, na* | *b'a, ni* | 0.238 | 0.036 |
| Terminal Classic | *b'a* | *a, ka, mi, na, u* | 0.228 | 0.028 |

Table 5.5: Periods in which syllables exhibit their highest and lowest intra-class average variabilities. The table provides also, the average variability for each period along with its standard deviation.

| | Syllables in region with | | **Variability** | |
|---|---|---|---|---|
| **Region** | **highest** | **lowest** | **average** | **std** |
| Petén | *la, na* | *a, ka* | 0.251 | 0.039 |
| Motagua | *a, b'a, ni, u* | *mi* | 0.258 | 0.057 |
| Usumacinta | *ka* | *ni* | 0.349 | 0.028 |
| Yucatán | *mi* | *b'a, la, na, u* | 0.214 | 0.033 |

Table 5.6: Regions for which syllables exhibit their highest and lowest intra-class average variabilities. The table provides also, the average variability for each region along with its standard deviation.

the average variation score and corresponding standard deviation for each historical period.

Analyzing the intra-class variability over time, we observed that 5 of the 8 classes reached their highest degree of visual variability in the Early Classic (200 - 500 AD), whereas only 1 syllable did it by the Terminal Classic (800 - 950 AD). Conversely, we found that only 1 class reached its lowest visual variability in the Early Classic, while 5 of them did it in the Terminal Classic.

Although the dataset is clearly small, and no strong conclusions should be made, after discussing with our archaeology partner, this analysis found some numerical differences over historical periods, and might (albeit weakly) suggest that the visual representation of the 8 syllabic classes went through a stabilization process across time (i.e., on average the variability goes from 0.277 to 0.238 and then to 0.228, also with decreasing standard deviation). Specific examples are syllables *a* (T229), *ka* (T25), *mi* (T173) and *u* (T1) whose variability decreases in subsequent periods. This is an open issue that would clearly require future work.

### Analysis over geographic regions

Table 5.6 shows the average variability across each of the regions. Region 2 (Motagua) seems to be the most variable as half of the classes reached their highest variability on this region, this observations is consistent with average variability score for this regions, which is the second highest value (0.258). On the other hand, Yucatán appears to be the less diverse region as half of the classes reached their lowest variability there. This is also reflected in its low average variability score (0.214).

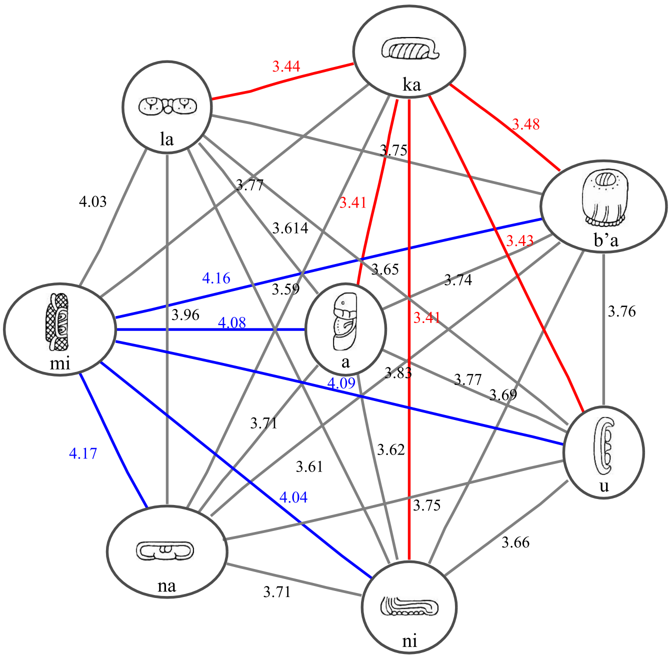


Figure 5.7: Inter-class similarity graph. Every node contains an example of the corresponding class, every edge is weighted with the similarity between the two nodes it connects to. The edges are colored to indicate similarity level; red (very different glyphs, smaller values), blue (very similar glyphs, higher values), and gray (middle level).

### Inter-class analysis

We define the distance between classes A and B to be the average of all the distances from each instance of class A to each instance of class B. We use the inverse of this distance as the link strength to form the graph shown in Figure 5.7.

The set of link strengths varies from 3.40 to 4.17, which implies that the relative difference between the most and the less similar pair of syllables is only of 18.3%. Interestingly, all the 5 less similarity edges (red in Figure 5.7) link the syllable *ka*, whereas the 5 most similarity ones (color blue) all connect the syllable *mi*. Respectively, syllables *ka* and *mi* are the least and most consistent classes in terms of visual similarity.

## 5.7 Conclusions

In this chapter, we proposed and developed the Histogram Of Orientation Shape Context (HOOSC) descriptor to overcome drawbacks found when using the Generalized Shape Context (GSC) [Mori et al., 2005]. The HOOSC is more robust as it describes the vicinity of a given pivot-point (within a shape) in a local fashion, excluding potential noise that could be present in the farther regions of the shape. It also improves the description itself by the incorporation of the distribution of local orientations rather that using only the dominant orientations.

In order to validate the HOOSC descriptor in Maya hieroglyphic data, we compiled a new digital set of syllabic Maya hieroglyphs that comprises 3400+ instances distributed over 24 positive and 1 negative classes. This dataset is unique in nature and highly valuable for the archaeology scholarly community, as it presents several challenges for visual description. To the best of our knowledge is the largest dataset of Maya hieroglyphs that has been analyzed with automatic tools.

We compared the descriptive performance of the GSC and the HOOSC in a set of image retrieval experiments performed under the *bov* approach. By using the HOOSC descriptor, we achieved 18.8% of absolute improvement in terms of retrieval precision. Overall, our results demonstrate that relevant elements are retrieved first for most of the cases, and that only a few of them fail, either because of their intra-class variability or because of the high visual similarity across some classes.

Also we organized a subset of the Maya syllabic dataset into synchronic and diachronic structures that allow to analyze the visual evolution of the Maya syllables along time and across geographical regions. Our results suggest that Maya glyphs tend to have less visual variability in subsequent periods suggesting a gradual visual convergence, observation that coincides with the previous conclusions of [Lacadena, 1995]. However, as glyphs started to disseminate across different regions, they were enriched with new visual features that increased their variability as the analysis across regions showed. In addition, we designed a similarity graph to visualize the visual similarity of Maya syllabic classes, which we trust can help epigraphers analyze visual relationships among classes of hieroglyphs.

Overall, we believe that the proposed descriptor is suitable for general shapes and that it will be able to handle other datasets. We think that our results will motivate the implementation of several systems to support queries of scholars in archaeology and, in the long term, from general audiences like visitors to museums.

# 6 Efficient *bov* representations for glyph retrieval

Sparse coding is a methodology based on observations from research work by the neuroscience community, which suggests that the receptive field on the mammalian primary visual cortex encodes natural images as sparse signals [Olshausen and Field, 1997]. The intuition behind sparse coding is that by representing an input signal as a sparse linear combination of certain set of bases called dictionary, the resulting vector will use only a small set of patterns (i.e., dimensions). Thus, sparse coding is expected to produce representations with high discriminative power.

In this chapter we use sparse coding as quantization technique to build *bov* representation of Maya hieroglyphs using HOOSC descriptors. We compare the sparse coding method with the *k*-means algorithm, and evaluate their performance in the tasks of retrieval of shapes.

Sparse coding has become common in a wide number of problems in computer vision and multimedia research, for instance, images and video inpainting and denoising [Mairal et al., 2008], image compression [Ranzato and LeCun, 2007], image restoration [Mairal et al., 2010], image classification [Boureau et al., 2010a,b], and shape representation [Mendels et al., 2006]. However, recent attempts to classify images based on sparse representations [Rigamonti et al., 2011] suggest caution using this technique, as perhaps it is not completely suitable to deal with non-natural images, or at least not when the level of noise is considerably high. A number of algorithms have been proposed for the tasks of learning dictionaries and estimating sparse decompositions [Lee et al., 2007]. Among them, K-Singular Value Decomposition (KSVD) was proposed as a generalization of *k*-means. This methods is easy to implement and has achieved good performance dealing with the image denoising problem [Aharon et al., 2006].

This chapter is organized as follows. Section 6.1 introduces the general theory of sparse coding. Section 6.2 explains the KSVD algorithm. Section 6.3 explains different pooling strategies to build bag representations. Section 6.4 explains our experimental protocol. Section 6.5 discusses our results. And section 6.6 presents our conclusions. This work has been reported in [Roman-Rangel et al., 2012].

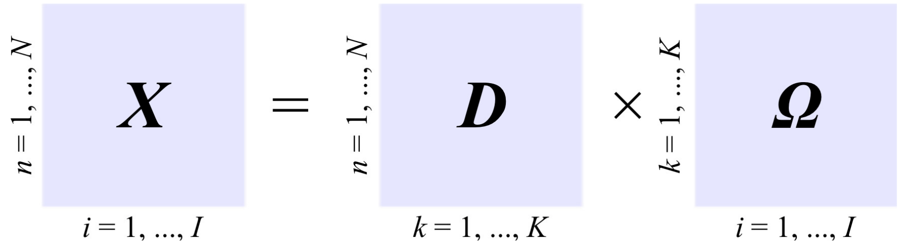


Figure 6.1: Visual representation of the sparse decomposition. The input $X$ has ($N$-dimensional) $I$ input signals, the dictionary $D$ has ($N$-dimensional) $K$ bases, and the resulting weighting matrix $\Omega$ has ($K$-dimensional) $I$ weighting vectors.

## 6.1 Sparse Coding

Sparse Coding is a method to decompose a set $X$ of signals, into sparse linear combinations of a set of basis functions called *dictionary*, and denoted by $D$ [Olshausen and Field, 1996]. In practice, the sparse representation of an individual signal $x_i$ correspond to the solution of

$$\min_{\omega} \|\omega\|_0 \qquad \text{s.t.} \quad x_i = D\omega_i, \tag{6.1}$$

where $\|\cdot\|_0$ denotes the $l_0$ pseudo-norm counting the number of non-zero entries in $\omega_i$, and $\omega_i$ is the weighting vector for the linear combination of the bases in the dictionary. Given that small errors might occur in the reconstruction of the original signal, the decomposition is often approximated as,

$$\min_{\omega} \|\omega\|_0 \qquad \text{s.t.} \quad \|x_i - D\omega_i\|_2 \leq \epsilon, \tag{6.2}$$

where $\|\cdot\|_2$ denotes the $l_2$ norm, and $\epsilon$ is an acceptable reconstruction error.

Usually, over-complete dictionaries are preferred in order to achieve a robust reconstruction of of the original signal. That is, a dictionary where the number of basis functions is much larger than the dimensionality of the input space [Aharon et al., 2006].

In section 6.2 we provide further details of sparse coding, and introduce the KSVD algorithm that has proven successful for sparse representations and image denoising.

## 6.2 KSVD

The work in [Tropp, 2004] presents sparse coding as a generalization of the quantization problem, consisting in the representation of the set of $I$ input signals $x_i$ as (sparse) linear combinations of the $K$ bases $d_k$ in the dictionary $D$, where the input signals are grouped as the columns of the matrix $X$, the bases are grouped as the columns of the dictionary matrix $D = [d_1, d_2, \ldots, d_K]$, and the resulting combination weights $\omega_i$ are grouped as the columns of the matrix $\Omega$, as shown in Figure 6.1.

Mathematically, the solution to this problem is approximated by,

$$\min_{D,\Omega}\{\|X - D\Omega\|_F^2\} \qquad \text{s.t.} \quad \forall i, \|\omega_i\|_0 \leq T, \tag{6.3}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm, $\|\omega_i\|_0$ is the $l_0$ pseudo-norm defining the number of non-zero entries in $\omega_i$, and $T$ is a parameter to control the number of basis functions allowed to be combined for the reconstruction of the input signals. In other words the constraint in Equation (6.3) allows $\omega_i$ to be a weighting row vector with more than one non-zero entry, such that matrix $\Omega$ has one weighting row $\omega_i$ for each input signal $x_i$. Note that Equation (6.3), and in general the sparse coding formulation, resembles the quantization problem presented in Equation (5.6).

The KSVD algorithm [Aharon et al., 2006] solves this minimization problem in two iterative steps. First, given a fixed dictionary $D$, the weighting coefficients $\Omega$, needed for the sparse decompositions, are found by the use of any pursuit algorithm like Matching Pursuit [Mallat and Zhang, 1993] or Orthogonal Matching Pursuit [Tropp, 2004]. After that, the dictionary is updated one basis at a time using singular value decomposition (SVD). The update of each basis $d_k$ is performed allowing changes in the components of the coefficients $\omega_i$ associated to it, which results in accelerated convergence [Aharon et al., 2006]. As in the $k$-means algorithm, a common option to choose the initial dictionary is to use a randomly selected subset of the input signals. Note that in order to achieve robust reconstructions of the input signals, the set of bases must be an over-complete dictionary, i.e., the number $K$ of basis functions must be (much) larger than the number of dimensions in the input signal space, $K \gg N$ in Figure 6.1.

### 6.2.1 KSVD with the $l_1$ norm

Based on the observation that the $l_1$ distance improves the results obtained by $k$-means in retrieval experiments [Roman-Rangel et al., 2011b], we evaluated the effects of combining it with the KSVD algorithm. More specifically, we changed the norm of the reconstruction error function presented in Equation (6.3) to be the $l_1$ norm. Thus for each input signal, we minimize:

$$\|x_i - D\omega_i\|_1 \text{ s.t. } \forall i, \|\omega_i\|_0 \leq T. \tag{6.4}$$

Note that KSVD relies in a singular value decomposition step which requires a $l_2$ normalization of the dictionary elements. We have not modified this normalization but only the reconstruction error function.

### 6.2.2 Sparse coding for HOOSC descriptors

With a slight abuse of terminology, a signal is said to be sparse if it can be decomposed into a sparse linear combination of a set of basis functions. We investigated the effects of enforcing

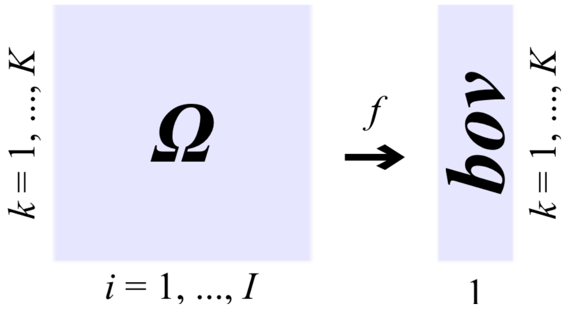


Figure 6.2: Schematic representation of the mapping from sparse coefficients to a *bov* representation for a given set of signals. In this example Ω has the weights that map $I$ signals into $K$ basis functions. Thus, the goal is to estimate a *bov* over the $K$ basis functions (or visual terms).

sparsity in the HOOSC descriptors by applying a threshold filtering to its components, as the initial set of experiments with the non-sparse HOOSC descriptor resulted in very poor retrieval precision.

More specifically, we set to zero all the components in the HOOSC descriptors whose value is below a certain threshold $\tau$. This thresholding step increases the sparsity of the input signals and facilitates their sparse decomposition. We observed that in practice, this step helped improve the average retrieval precision.

## 6.3 Building bag models from sparse coefficients

When the quantization is made via $k$-means, each descriptor is associated to a single cluster, such that computing the *bov* representation for a given image is as simple as counting how many instances of each visual word this image has (i.e., descriptors of each cluster). However, the sparse coding approach associates each descriptor to several bases (*visual words*), where the (sparse) coefficients denote the strength of that association. Thus, we can explore different pooling criteria to find a function $f$ that maps the sparse coefficients into *bov* vectors.

Figure 6.2 shows a schematic representation of the process of mapping the matrix Ω that contains the sparse coefficients of a given set of descriptors (hieroglyph) to its *bov* representation. Note that the matrix Ω is indexed by $k = 1,\ldots,K$ (number of bases), and $i = 1,\ldots,I$ (number of local descriptors of the given glyph); whereas the *bov* is a vector indexed only by $k = 1,\ldots,K$.

Some of the pooling approaches available to compute *bov* representations are:

- Average Pooling (AVP). For a given glyph, it assigns to each visual word the average of its corresponding responses computed over the whole set of descriptors. In other words, the final *bov* is the average of the absolute values of each row in Ω,

$$\tilde{bov}_k = \frac{\sum_i abs(\Omega_{ki})}{I}, \tag{6.5}$$

where $abs(\cdot)$ denotes absolute value, and $I$ is the number of local descriptors for the given glyph. The need for using the absolute values is because some of the weight might have negative responses. We are interested in the strength of the response, regardless of its direction.

- Max-K Weight Pooling (Max-KWP). It consists in building the *bov* vector as the sum of the weights of the $K_{max}$ coefficients having the maximum responses, where $K_{max}$ is kept constant for all the weighting vectors. More precisely, let $f_{K_{max}} : \Omega \rightarrow \Omega^{K_{max}}$ be the mapping that generates a copy of $\Omega$ setting all its entries to zero, except for those corresponding to the $K_{max}$ components in each column of $\Omega$, i.e., for each vector $\omega_i$, $f_{K_{max}}$ keeps only the $K_{max}$ maximum responses. The *bov* is then computed as

$$\tilde{bov}_k = \frac{\sum_i abs\left(\Omega_{ki}^{K_{max}}\right)}{K_{max}}. \tag{6.6}$$

- Max-K Binary Pooling (Max-KBP). It builds the *bov* representation as the binary activation of the basis function associated with the coefficients having the maximum responses,

$$\tilde{bov}_k = \begin{cases} 1 & \text{if} \sum_i \left| abs\left(\Omega_{ki}^{K_{max}}\right) > 0 \right| > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{6.7}$$

where $|\cdot|$ denotes cardinality, i.e., the number of times its argument becomes true. In other words, a basis is activated in the *bov* as soon as it is a significant basis for at least one of the local descriptors of the glyph.

- Max-K Integer Pooling (Max-KIP). This approach builds the bag representation as the integer count of the basis function associated with the $K_{max}$ coefficients having the maximum responses. The *bov* representation is computed as

$$\tilde{bov}_k = \sum_i \left| abs\left(\Omega_{ki}^{K_{max}}\right) > 0 \right|, \tag{6.8}$$

Note finally, for each of the above possible schemes, the *bov* representations are normalized vectors, i.e., all of the above mentioned vectors are normalized as,

$$bov_k = \frac{\tilde{bov}_k}{\sum_{j=1}^{K} \tilde{bov}_j}. \tag{6.9}$$

Among them, the *Max-1 Integer Pooling* (Max-1IP) method seems to have given the best performance in previous works for image classification [Boureau et al., 2010b].

### 6.3.1 Building process summary of bag models from sparse coefficients

In summary, the process to build *bov* representations using sparse coding is:

1. Compute local descriptors for each of the Maya hieroglyphs in the dataset.
2. Use a subset of local descriptors as input to KSVD (or other dictionary learning technique), and estimate a dictionary with basis functions.
3. Compute the sparse decomposition of all the local descriptors for all the input images (Maya glyphs).
4. Use one of the pooling schemes to construct a *bov* representation for each glyph.

In the following section, we present the experimental protocol we followed to evaluate the use of sparse coding in the task of retrieving Maya glyphs.

## 6.4 Experiments

In this section we explain the protocol followed during the evaluation conducted to assess the performance of the sparse coding and clustering approaches in the construction of *bov* representations of shape images. We also comment the data used.

### 6.4.1 Data

For all the experiments in this chapter we used the syllabic Maya dataset (SM) introduced in section 5.4.1, which consists of 1270 syllabic Maya hieroglyphs that are distributed over 24 visual classes. All the 24 classes are subdivided in two subsets: *candidates* ($G_C$) and *queries* ($G_Q$). Around 80% of examples of each class are selected as candidates and used to build the representation model (clusters with $k$-means, or sparse dictionary with KSVD). The remaining 20% of the examples of each class are used as queries to evaluate the retrieval performance of the studied indexing techniques.

### 6.4.2 Experimental protocol

We compared three different formulations of the HOOSC descriptor:

1. HOOSC: corresponds to the HOOSC as it is presented in [Roman-Rangel et al., 2011a], which uses the intermediate spatial context of the polar grid, i.e., rings 2, 3, and 4, resulting in a vector of 290 dimensions.
2. sHOOSC: is the HOOSC with only 194 dimensions, as explained in chapter 5.

3. stHOOSC: is the same as sHOOSC after been thresholded as explained in section 6.2.2.

We started performing the dictionary learning process via $k$-means or KSVD. From the *candidates* subset ($G_C$), we chose randomly 1500 descriptors from each of the 24 classes, thus 36000 in total, and used them to estimate dictionaries of different sizes. Then, using the dictionary model of ($G_C$), we computed the *bov* representation for all the glyphs. For the KSVD cases, the *bov* construction step was repeated several times according to different pooling techniques.

As we have mentioned, we implemented the combination of the $l_1$ distance with KSVD (see section 6.2.2). In order to compare the reconstruction error between $k$-means and KSVD, we used both $l_1$ and $l_2$ distances with the sparse coding approaches. We refer to the possible combinations as KSVD-$l_1$ and KSVD-$l_2$, and $k$-means-$l_1$ and $k$-means-$l_2$, respectively.

During the retrieval experiments, we build the *bov* representation of each query glyph using each of the tested pooling techniques, and compare it against the *bov* of the glyphs in $G_C$ using the $l_1$ distance, then we ranked the resulting distances. Finally, we estimate the Average Precision (*AP*) of each query from the ranking of the candidate glyphs belonging to the query class, and compute the *mAP* of the current representation to evaluate its performance.

In summary, the experiments we performed are:

1. We evaluated the retrieval performance of the four pooling techniques explained in section 6.3 to build the *bov* representations. To this end, we considered different numbers $K_{max}$ of the bases having the maximum responses (see section 6.5.1).

2. We evaluated the impact of the thresholding procedure used to enforce sparsity for different values of the parameter $\tau$ (see section 6.5.2).

3. We performed retrieval experiments using dictionaries of different sizes, estimated with $k$-means-$l_1$, KSVD-$l_1$, and KSVD-$l_2$ (see section 6.5.3). Empirically, we saw that $k$-means-$l_2$ does not improve the retrieval precision over $k$-means-$l_1$. We do not show those results here.

4. To acquire a clearer idea of the behavior of the mean Average Precision (*mAP*) of the different approaches, we compared the reconstruction errors achieved by $k$-means and KSVD, both with $l_1$ and $l_2$ distances (see section 6.5.5).

5. To investigate the combination of methodologies that better discriminate visual classes of glyphs, we have computed the inter-class distance between two syllabic classes $A$ and $B$ as the average pair-wise distance between each instance of class $A$ with respect to each instance of class $B$ (also obtaining the intra-class distance when $A = B$). We performed this inter-class similarity study comparing the $k$-means and KSVD approaches that achieved the best retrieval precision, and using two different distance metrics: Euclidean and Jensen-Shannon Divergence [Lin, 1991] (see section 6.5.7).

6. We did a study to evaluate the potential of our methods to automatically discover visual patterns in shape descriptors of Maya hieroglyphs. A tool with such a capacity is of great interest for archaeologists, as it could suggest visual similarities of symbols based on local visual patterns. To this end, we localized the most frequent visual words in each class and its associated closest pivots, then we looked at its neighboring points that are used to construct its HOOSC descriptor (see section 6.5.8).

## 6.5 Results

In this section we present the results of our extensive evaluation. Note that the combinatorial nature of our experiments produced multiple results. Therefore, we only discuss those results that are relevant to each subsection to facilitate their reading.

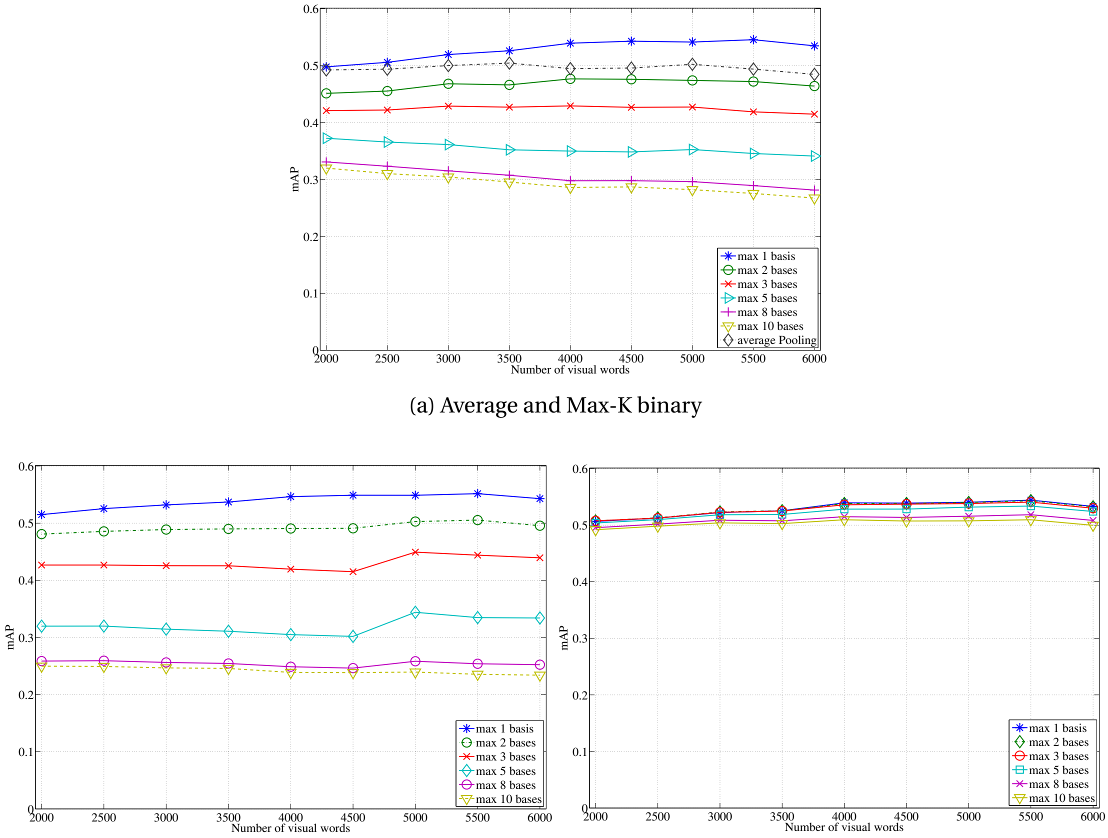
### 6.5.1 Pooling scheme evaluation

First, we evaluated the performance of the four different pooling schemes to construct *bov* representations based on KSVD: Average Pooling (AVP), Max-K Binary Pooling (Max-KBP), Max-K Integer Pooling (Max-KIP), and Max-K Weighted Pooling (Max-KWP). Figure 6.3a shows the *mAP* retrieval results obtained when comparing *bov* vectors that are computed using the AVP (dashed-diamond line), it also shows the performance curves for the Max-K Binary Pooling for various values of $K_{max}$. In general, using hard assignments to only the basis with the highest response gives better results than any of the other options.

In Figure 6.3b, the retrieval results of Max-KIP are shown for distinct values of $K_{max}$. Note that the curves in the binary and integer cases have similar behavior, and that the less coefficients used to estimate the *bov* representation, the better the retrieval results. Since the coefficients represent weights for linear combinations, this might sound counter intuitive as it could be expected that a weighted assignment to visual words could help reconstruct better the original signal. This was not the case in practice.

The curves shown in Figure 6.3c present results when the *bov* are computed combining the actual weights of the coefficients corresponding to the highest responses (Max-KWP). In this case, varying the maximum number of coefficients has little impact in the performance. Also, all of the results in Figure 6.3c perform below the Max-1 Integer pooling shown in blue-diamond in Figure 6.3b. Furthermore, the Max-1 Integer pooling strategy outperforms any of the other approaches.

Looking at the number of visual words, the results show that this parameter has little impact in the retrieval performance.

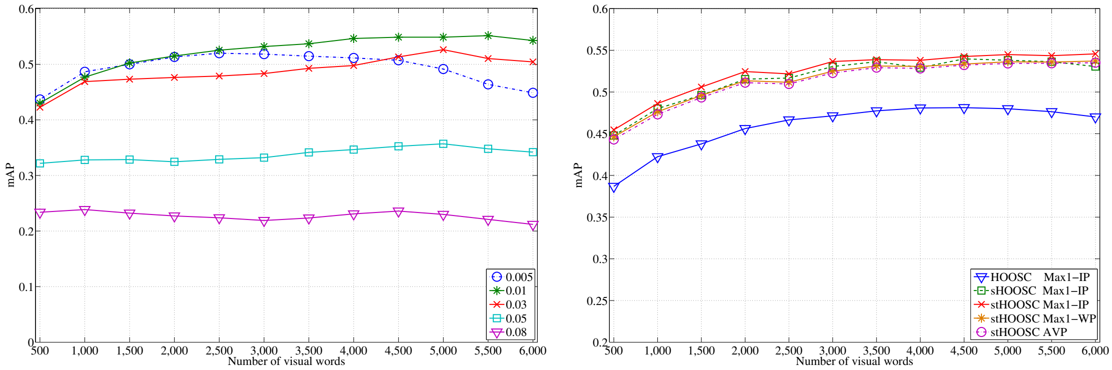


(a) Average and Max-K binary

(b) Max-K integer

(c) Max-K weighted

Figure 6.3: Retrieval precision (*mAP*) of different pooling techniques to compute *bov*'s from sparse coefficients computed with 'euclidean' distance.



(a) *mAP* for several threshold values ($l_2$ distance).

(b) *mAP* using the $l_1$ distance.

Figure 6.4: Retrieval results with KSVD and different number of bases in the dictionary.

### 6.5.2 Facilitating the sparse decomposition

We compared the performance of the KSVD method to build *bov* representations after performing a thresholding step that sets to zero all the HOOSC components below the threshold $\tau$. By doing so, the sparse decomposition of the HOOSC descriptors results in *bov* representations that allow for better retrieval precision. Figure 6.4a shows these results. We can see that $\tau = 0.01$ provides slightly better result than $\tau = 0.005$ and $\tau = 0.03$, and that higher threshold values generate very poor results.

Nota that opposite to the curves shown in Figure 6.3, the use of the thresholding procedure results in a method with higher response to changes in the vocabulary size. For instance, the curve in in Figure 6.4a corresponding to $\tau = 0.01$, shows a difference of a rough 10% improvement when using 500 or 4500 bases.

### 6.5.3 Combining $L_1$ with KSVD

The combination of the KSVD algorithm with the $l_1$ (city-block) distance (KSVD-$l_1$) resulted in a slight decrease of the retrieval precision with respect to the original KSVD that uses the $l_2$ distance. Figure 6.4b presents the subset of the most relevant curves resulting from this assessment.

We noticed that these curves behave similarly when we vary the value of the threshold $\tau$, and that there is not significant difference when the pooling strategy is changed. Regarding the HOOSC formulation to be used, there is an important improvement in the retrieval precision achieved when the most central spatial context is merged into a single *ring* (sHOOSC), with respect to the version presented in [Roman-Rangel et al., 2011a] (HOOSC). However, the improvement after performing the thresholding procedure (stHOOSC) remains modest.

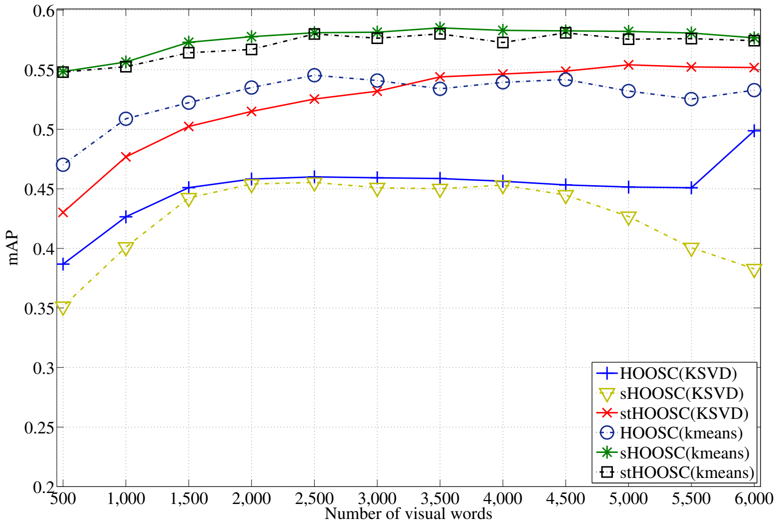


Figure 6.5: *mAP* for the different versions of the HOOSC with varying number of visual words: HOOSC is the method in [Roman-Rangel et al., 2011a]; sHOOSC and stHOOSC are respectively the method explained in chapter 5 before and after performing the thresholding.

### 6.5.4 Comparing HOOSC formulations

All the results discussed through sections 6.5.1, 6.5.2, and 6.5.3 were computed for all possible combinations of "pooling strategy - threshold value - HOOSC formulation", thus resulting in a large amount of tables to be analyzed. As already mentioned, we only present the most relevant results to facilitate their reading, i.e., the results shown about **pooling schemes** correspond to those computed with the threshold fixed to $\tau = 0.01$, whereas the curves regarding the **facilitation of the sparse decompositions** correspond to using Max-1 Integer pooling. In all cases we present the results obtained with the HOOSC with 194-D explained in chapter 5.

Now, we discuss the impact in the retrieval precision obtained for the different constructions of the HOOSC descriptor. In Figure 6.5, we compare the performance obtained when using the KSVD approach with the best combination of parameters, i.e., with the $l_2$ distance (KSVD-$l_2$), using Max-1 Integer pooling, and using a threshold fixed to $\tau = 0.01$.

Different from the result obtained with KSVD-$l_1$, these results show that the simplified HOOSC (sHOOSC) performs slightly lower than the original HOOSC with a drastic decrease after 4500 bases when the KSVD method is applied. However, after the implementation of the thresholding procedure (stHOOSC), its performance is notably increased, reaching its maximum when 5000 bases are used as dictionary elements.

Figure 6.5 also shows the results obtained by the $k$-means quantization method (for these experiments we only show the results of the 'city-block' distance). The simplified HOOSC (sHOOSC) has a consistently better performance, around 5% more than the HOOSC in [Roman-Rangel et al., 2011a]. Note that the performance of the descriptor that uses the threshold step

| **visual-words** | 1000 | 2000 | 3000 | 4000 | 5000 | 6000 |
|---|---|---|---|---|---|---|
| $k$-means-$l_1$ | 1.233 | 1.127 | 1.029 | 0.976 | 0.914 | 0.861 |
| KSVD-$l_1$ | 1.452 | 1.413 | 1.363 | 1.374 | 1.440 | 1.519 |

Table 6.1: Quantization error of clustering and sparse coding with the $l_1$ distance for different number of visual words.

(stHOOSC) has no considerable difference with respect to the simplified HOOSC. In general, the performance of the three versions of the descriptor tends to degrade after 4000 clusters.

Overall, we noticed that KSVD does not seem to achieve as good retrieval results as the traditional $k$-means method. The best result obtained with KSVD (0.554) using 5000 bases and the euclidean distance, is lower than the corresponding result of 5000 clusters of $k$-means (0.582), and lower than the best result of $k$-means, obtained with only 3500 clusters (0.585).

### 6.5.5 Comparing the quantization error

To better understand the behavior of $k$-means and KSVD as quantization approaches, we compare the reconstruction error achieved by both methods. Note however that it is difficult to compare the best performing method for KSVD (which is KSVD-$l_2$) and $k$-means (which is $k$-means-$l_1$), as they use different metrics. By nature these two metrics can have different order of magnitude and, a direct comparison of the different reconstruction errors might not be correct. This is due to that the $l_1$ distance accumulates the absolute sum of the dimension-wise differences between two vectors, while the $l_2$ correspond the their euclidean distance, i.e., in the case of density functions, $l_1$ will result in higher values. To address this issue we compared methods when using the same reconstruction metric.

In Table 6.1, we show the average reconstruction error achieved by $k$-means and KSVD during the dictionary learning process using the $l_1$ distance. In each case, we present the result corresponding to the best HOOSC formulation, that is: sHOOC for $k$-means and stHOOSC for KSVD, with 3500 and 5000 visual words respectively.

We can see that $k$-means has a consistent lower reconstruction error than KSVD when the $l_1$ distance is used. Also note that, for $k$-means, the reconstruction error tends to decrease as the number of bases increases, whereas it exhibits a local minimum at 3000 bases in the case of KSVD.

In Table 6.2 we show similar results computed with the $l_2$ distance. In this case the reconstruction error continues decreasing as the number of bases increases for both algorithms, and $k$-means remains consistently as the method with lower reconstruction error.

| **visual-words** | 1000 | 2000 | 3000 | 4000 | 5000 | 6000 |
|---|---|---|---|---|---|---|
| $k$-means-$l_2$ | 0.032 | 0.030 | 0.026 | 0.023 | 0.018 | 0.012 |
| KSVD-$l_2$ | 0.270 | 0.267 | 0.201 | 0.145 | 0.132 | 0.127 |

Table 6.2: Quantization error of clustering and sparse coding with the $l_2$ distance for different number of visual words.

| | $k$-means | KSVD |
|---|---|---|
| Euclidean | 75.0 | 75.0 |
| $D_{JS}$ | 100.0 | 100.0 |

Table 6.3: Percentage of times the intra-class distance is minimal compared to the inter-class distance.

### 6.5.6 Visual result comparison

To graphically illustrate our results, in each row of Figure 6.6 we present one retrieval example per class. The first column shows the corresponding best query in each class, i.e., the query with the highest *AP* value. The remaining columns correspond to the most similar *candidate* ($G_C$) hieroglyphs retrieved in the top 10 positions of the ranking vector. The candidates enclosed in a blue rectangle correspond to relevant documents (hieroglyphs of the same visual class). These results have been generated using the best retrieval method, i.e., $k$-means-$l_1$ with sHOOSC (Figure 6.5). Note that some classes are easy to match as most of the top 10 retrieved elements proved to be relevant. However, there are still few classes whose elements are confused due to a high inter-class visual similarity, e.g., some classes share visual patterns such as lattices and horizontally elongated shapes.

### 6.5.7 Distance estimation between visual classes

As mentioned in section 6.4, having a method to estimate class distances of Maya hieroglyphs would allow to find the most probable visual classes of new discovered symbols. Figure 6.7 shows the inter-class distances for each pair of classes, computed as Euclidean distance and as the Jensen-Shannon divergence ($D_{JS}$) for the best results of $k$-means-$l_1$ and KSVD-$l_2$, i.e., 3500 clusters and 5000 bases, respectively.

In general, using $D_{JS}$ the intra-class distance is always smaller than the inter-class distances, whereas this is not true for the Euclidean distance. This suggests that by using $D_{JS}$, it might be possible to discriminate new symbols with higher accuracy. Table 6.3 shows the percentage of the average number of times the intra-class distance is smaller than the inter-class distance.

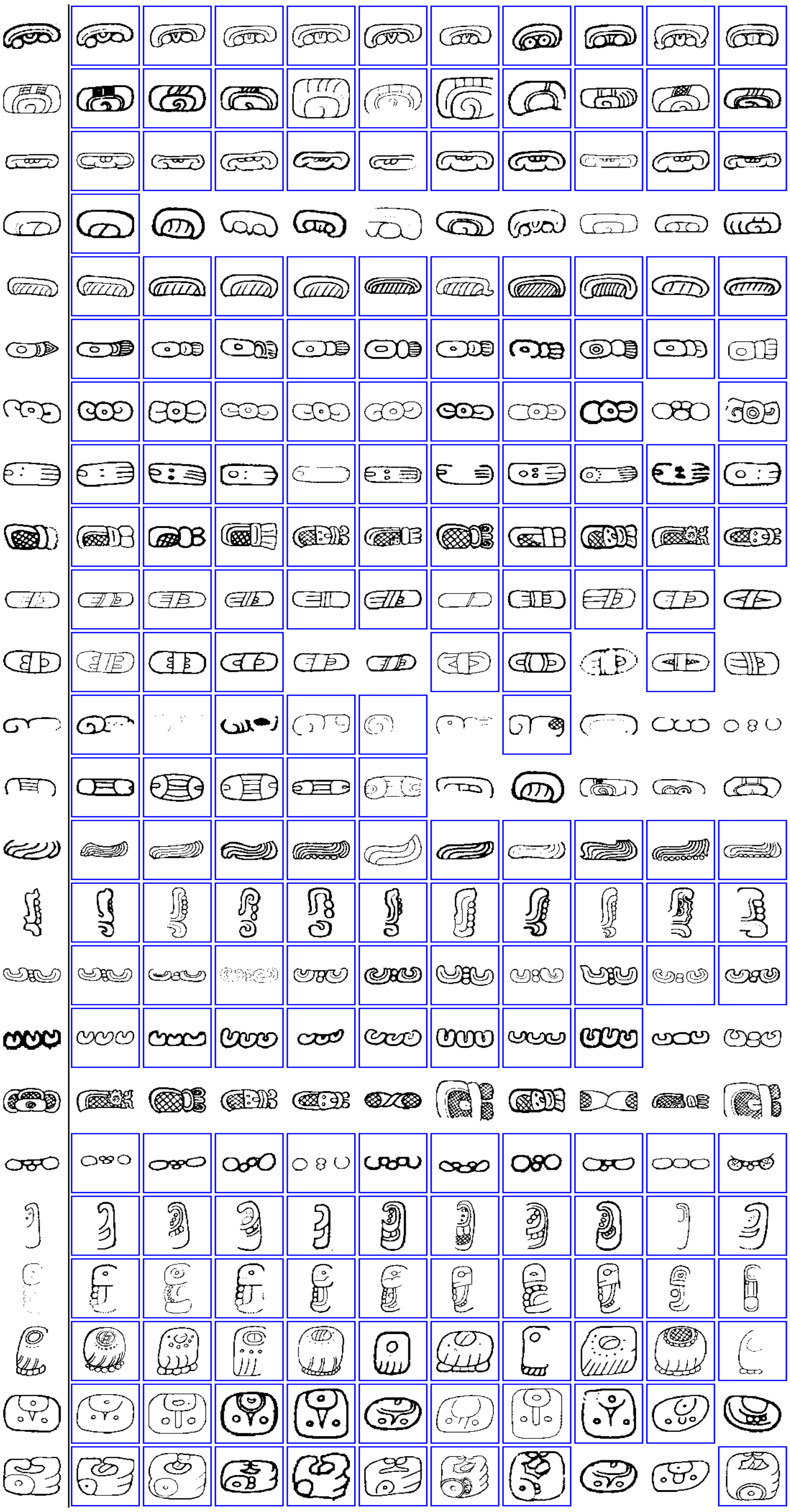
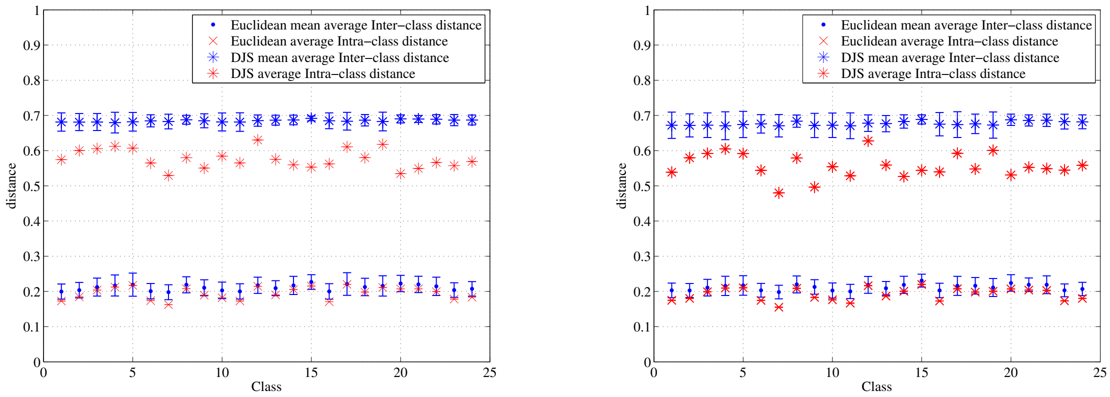


Figure 6.6: One retrieval example per class: the first column has one query per visual class, the remaining columns correspond to the top 10 retrieved hieroglyphs for each query. Relevant glyphs are enclosed in a blue rectangle. © AJIMAYA.



(a) *k*-means

(b) KSVD

Figure 6.7: Intra-class and Inter-class distances computed with Euclidean distance and Jensen-Shannon divergence.

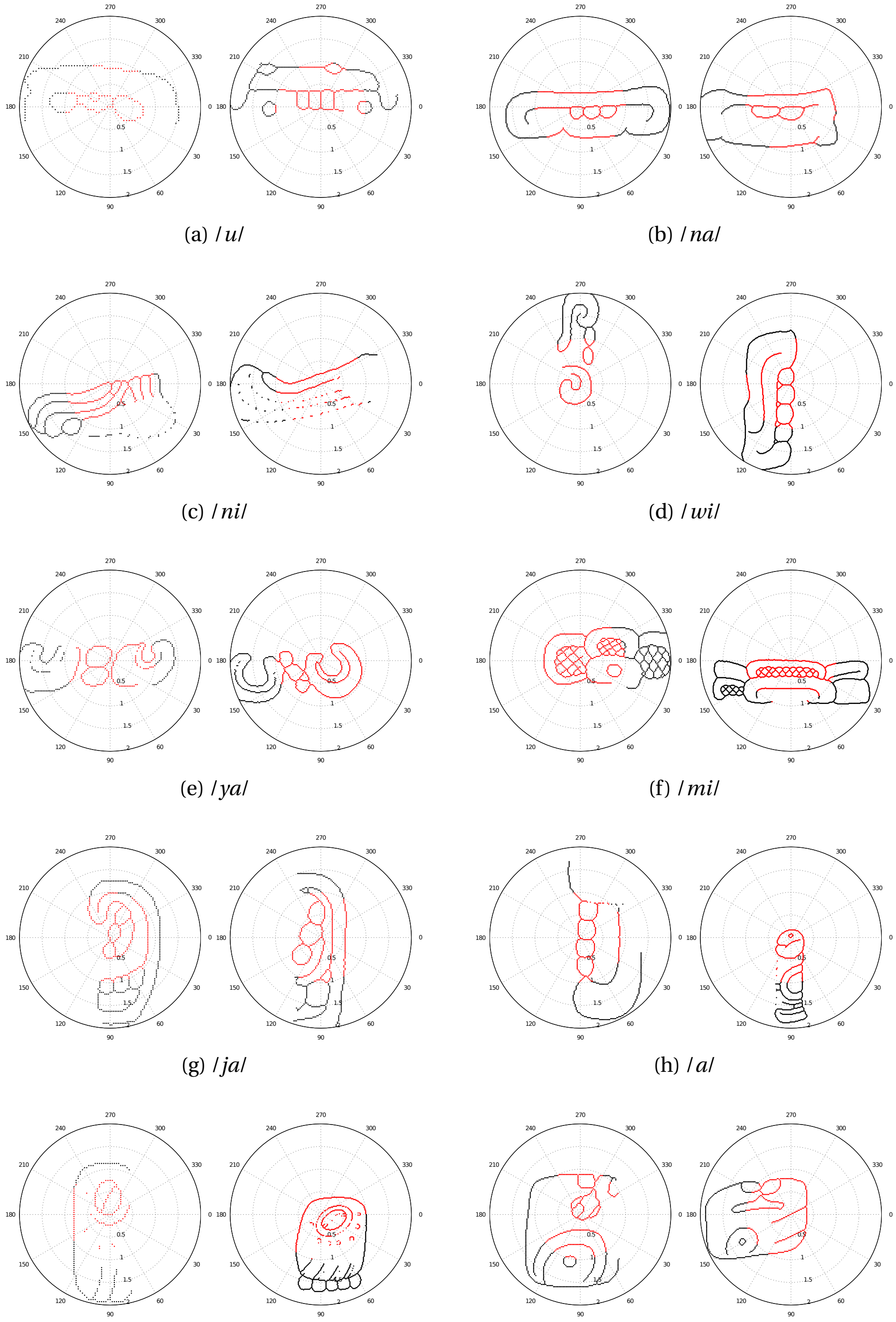
### 6.5.8 Visual pattern recovery

We noticed that some visual patterns (visual words) are more descriptive than others for certain classes, i.e., some visual patterns contribute more than others in the *bov* representations of glyphs within a given class. From the 3500 visual words estimated with *k*-means-$l_1$ and the sHOOSC approach, in Figure 6.8 we show graphical examples of the two most common visual words for some of the Maya syllabic classes. Each graphical example corresponds to the closest point to the two most popular clusters used in the *bov* representations within each class. Note that the sHOOSC method only uses the spatial scope up to one time the average distance of the set of points (see chapter 5), therefore only the red points in Figure 6.8 are part of the descriptors; we show the whole image with the purpose of providing visual context to the reader.

In general terms, we are able to retrieve *bov* representations that are visually consistent between glyphs of the same class, i.e., they use the same visual patterns in similar proportions. We believe that in the future, this observation might be useful to describe Maya hieroglyphs based on localized visual patterns automatically discovered.

## 6.6 Conclusions

The main contribution of this work is the evaluation of the performance of two quantization approaches in the construction of bag representations of local shape descriptors, and more specifically, *k*-means and KSVD. Sparse coding is a recent trend that has gained popularity for description and classification of natural images, and our work is a first exploration of the use of sparse coding decompositions as a quantization method of contextual shape descriptors. We have evaluated the retrieval performance of this method with different pooling strategies. We believe that this assessment allows to confirm the conclusions from previous classification



(a) /*u*/

(b) /*na*/

(c) /*ni*/

(d) /*wi*/

(e) /*ya*/

(f) /*mi*/

(g) /*ja*/

(h) /*a*/

(i) /*ba*/

(j) /*chi*/

Figure 6.8: The two most common visual patterns for visual class shown on two glyphs where they appear. This patterns are recovered by sHOOSC under $k$-means-$l_1$ clustering. The whole glyph is plotted to show visual context, though only the points in red are actually used for computing the description of the center pivot.

literature that state that (i) depending on the given application, sparse techniques might or not perform better, and (ii) in the context of sparse techniques that rely on pooling strategies, max-pooling often provides the best results.

We proposed an efficient method to facilitate the sparse decomposition of the HOOSC descriptor, this method consists in setting to zero all the signal components that are below a certain threshold. The evaluation of this method shows that the bag of visual words representations can achieve better retrieval performance if applied. We also compared the performance of KSVD with the traditional $k$-means clustering for dictionaries of different sizes, and found out that in general $k$-means performs better. This is an interesting results given the simplicity of $k$-means with respect to KSVD.

In our study, we also proposed a method to measure distances between pairs of visual classes of Maya hieroglyphs. This measure can be used to analyze visual relationships among glyph classes. Finally, we manually analyzed the visual patterns that our methods are able to encode and recover. We observed that the visual patterns encoded by HOOSC descriptors via $k$-means clustering are consistent across shapes that share similar visual aspects. We believe that this method could be potentially used to discover visual patterns that represent hieroglyphs, and shapes in general, in a robust manner.

# 7 Three applications of the HOOSC descriptor

In this chapter we present three applications of the HOOSC descriptor. Namely, the development of the first version of a content-based retrieval system for visual instances of Maya hieroglyphs, and an evaluation of the generalization of the HOOSC descriptor to deal with two datasets of shapes having instances of different nature to the Maya hieroglyphs.

In practice, we evaluated which of the HOOSC characteristics are suitable or beneficial to describe and retrieve shape data of different nature. We conducted experiments on two additional datasets: a set of ancient Chinese characters, and the MPEG-7 Core Experiment CE-Shape-1 test set [Latecki et al., 2000].

Part of this work was developed in collaboration with Microsoft Research Asia (MSRA) during a research internship in the Web Search and Mining group, having Dr. Changhu Wang as mentor, and has not been previously published. The rest of the content of this chapter is reported in [Gatica-Perez et al., 2011, Roman-Rangel et al., 2011a, 2012].

## 7.1 Towards a visual retrieval system

One of the long term objectives of our research is the implementation of an accurate visual retrieval system for Maya hieroglyphs. This tool will allow archaeologists to quickly search in large collections for instances of visual queries. The current version of this tool (in Matlab) has three modules:

1. **Retrieval of segments:** This module allows to query a glyph-instance that is manually selected from a set of instances that have been previously segmented and cleaned. The system computes the query description using HOOSC descriptors, and compares it agains a pre-indexed set of other segmented glyphs. Finally, it retrieves the $N$ most similar instances from the dataset (e.g., $N = 8$).
2. **Retrieval of selections:** This module works similar to the module **retrieval of segments**, except that the query is manually selected from a large inscription by marking a bound-

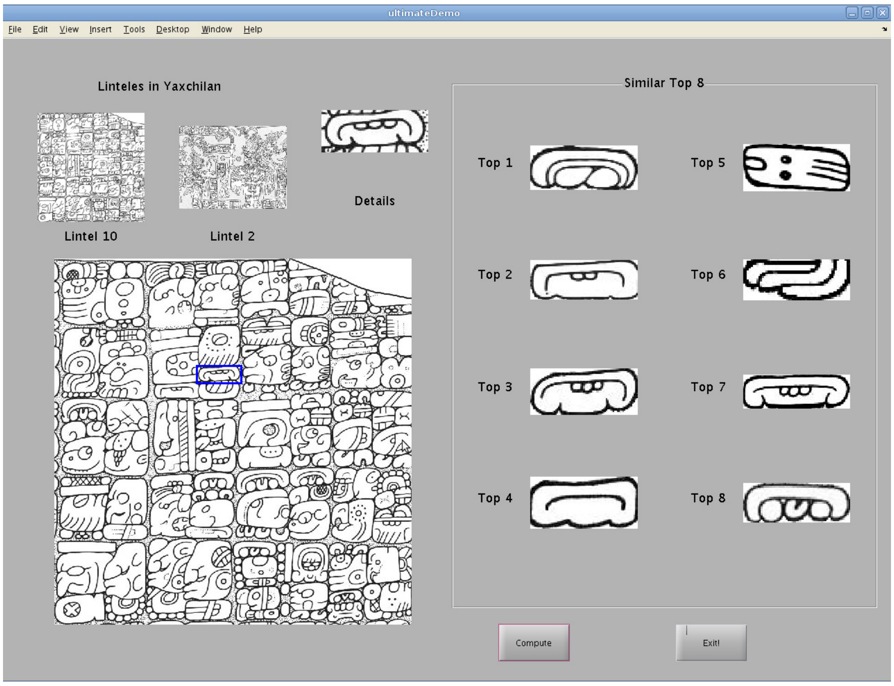


Figure 7.1: Snapshot of the demo version of a retrieval machine of Maya hieroglyphs.

ing box with the mouse. Therefore, the selected query might contains noise around it as it has not been previously cleaned.

3. **Detection of segments:** This module will perform the detection of segmented instances in large inscriptions. Currently it works as a demo, and it is ready to plug into it detection approaches, such as the one explored in chapter 8.

Figure 7.1 shows a snapshot of the second module retrieval of selections where the selected instance is highlighted in blue, and the $N = 8$ most similar pre-indexed glyphs have been retrieved from the dataset.

A video demo of this preliminary tool is available in the CODICES project website[1].

## 7.2 Chinese characters: The Oracle Bones Inscriptions

The Oracle-Bones Inscriptions (OBI) is a collection of ancient Chinese characters that have been found carved on animal bones and turtle shells, mainly nearby the city of Anyang, in the Henan province of China. This site corresponds to the last capital of the Shang dynasty. These bones are usually referred to as Oracle-Bones as they were used with divination purposes during the Bronze age of ancient China (starting c.a., 2000 BC) [Flad, 2008]. Figure 7.2 shows an Ox scapula from the Shang Dynasty (1600 BC – 1046 BC).

[1] http://www.idiap.ch/project/codices

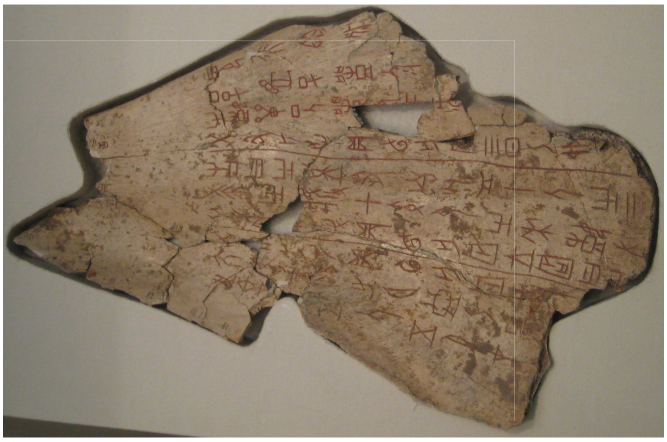


Figure 7.2: Ox scapula with divination inscriptions. ©Wikipedia.

Divination in ancient China was a common practice to predict the weather, the season for sowing, the outcome of trades and military endeavors, and fortune in general. The inscriptions in bones are shapes with rectilinear patterns, such as the examples shown in Figure 7.3. The patterns of these characters are different from those made by ink or in bronze, perhaps to make the carving process more efficient. Often, these characters were rotated to better fit the surface of the bone.

We performed retrieval experiments on a collection of segmented OBI characters that has been compiled by MSRA. This dataset consists of 31,784 segmented shape instances. From which, 24,100+ instances are distributed over 952 visual classes, and 7,600+ are not annotated. The number of instances per class in not uniform, thus the distribution of instances goes from classes with only one instance up to 291 instances in the most populated class, with more than 700 classes having 30 instances or less, and only 62 classes having 100 instances or more. Although the shapes in this dataset can be considered of similar visual complexity as the Maya hieroglyphs, the OBI dataset poses more challenges for description and retrieval, as it is much larger, and the number of instances per class has high variation. Also, the instances in the OBI dataset have not been manually aligned, i.e., some classes contain rotated and reflected instances.

### 7.2.1 Experimental settings

We used the SC and GSC descriptors as described in chapters 4 and 5, respectively, and compared their performance in retrieval tasks. We also used a variant of the HOOSC descriptor having the following characteristics: (1) uses thinned shapes as input; (2) computes descriptors at pivots with respect to point, however, we considered all the points as pivots as the shapes are relatively smaller in comparison with the Maya hieroglyphs, such that the computational efficiency is not affected by using all the points as pivots, and the description becomes more robust; (3) uses the 8-bin histogram of orientation per region; (4) uses two distance intervals

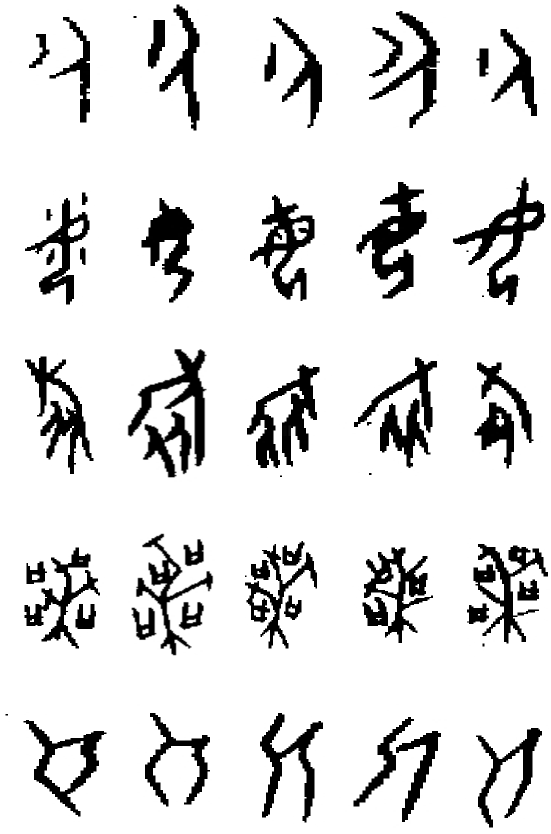


Figure 7.3: Examples of ancient Chines characters in the Oracle Bones Inscriptions. Each row shows a different visual class.

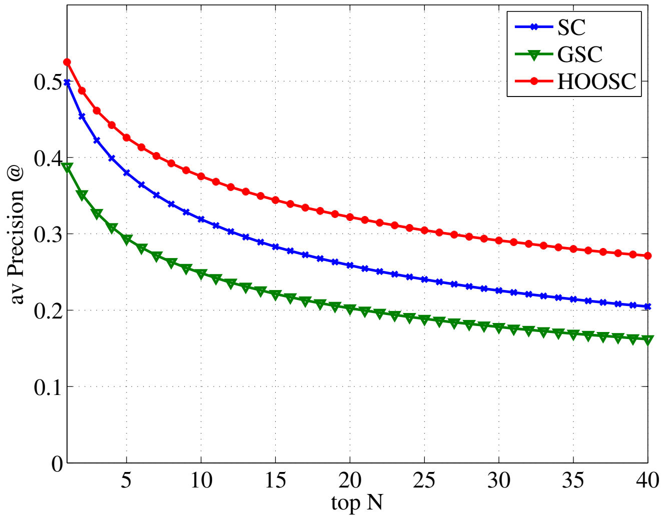


Figure 7.4: Average precision achieved with SC, GSC, and HOOSC on the OBI dataset.

with boundaries at 0.5 and 1.0 times the average pairwise distance among the points in the thinned contour; (5) as there are rotated instances, we did not use the relative self-position.

We provided the three descriptors with rotation invariance by aligning the local orientation of each point with the horizontal axis, and updating the computations accordingly. We used a *bov* approach with all three methods. Note that this is different from the SC framework introduced in chapter 4, where the comparison of shapes was done under a point-to-point matching approach. Using a *bov* approach the comparisons are much faster. We used a fixed number of 1000 visual words for all three descriptors. The visual vocabularies were computed selecting at random 1 descriptor per instance, and using the *k*-means clustering algorithm.

Since no supervised learning was used, we performed full cross-validation retrieval experiments, i.e., we used the 31,784 segmented instances as both queries and also elements to be retrieved from the collection. We compared the performance of the three methods by computing the retrieval precision at each of the first 40 instances ranked as most similar for those instances having a relevant set, i.e., characters from classes with at least two instances. We then computed the average of each of the 40 position to obtain a curve showing the performance of each method (avPrecision@top-N).

### 7.2.2 Retrieval results

Figure 7.4 shows the average precision achieved with each descriptor as the avPrecision@top-N curves. Similar to the results on the Maya syllables, the SC performs better than the GSC, and the HOOSC exhibits the curve with the best retrieval performance. Note that the difference in performance between SC and HOOSC becomes larger as more positions in the ranked vector are considered.

| | Precision@ | | Recall@ | |
|---|---|---|---|---|
| Descriptor | 1 | 10 | 1 | 10 |
| SC | 0.498 | 0.319 | 0.009 | 0.043 |
| GSC | 0.388 | 0.249 | 0.006 | 0.029 |
| HOOSC | **0.525** | **0.375** | 0.009 | 0.050 |

Table 7.1: Average retrieval precision and average recall at top 1 and top 10 for the different shape descriptors.

Table 7.1 summarizes the retrieval performance in terms of average precision and average recall at the 1st and 10th ranked elements. HOOSC shows the highest performance on average, however, the improvement here is only about 4% in absolute terms, thus not as large as in the case of Maya hieroglyphs. There are two reasons for this. First, we did not use the relative self-position in the descriptor, which in the case of the Maya glyphs accounts for a rough 3% improvement with respect to a HOOSC formulation that does not use it (see Tables 5.2 and 5.3 in section 5). Second, this dataset includes rotated instances, making more difficult to retrieve all the relevant instances at the first positions, whereas the syllabic Maya dataset has been manually aligned.

Finally, we show visual examples of the retrieved vectors. The first column of Figure 7.5 shows 12 images randomly selected from the OBI dataset. The remaining columns show the documents from the dataset that the version of HOOSC used in this chapter ranks as the most similar images, i.e., from top 1 to top 8 going from left to right. From the results, we can see that the HOOSC descriptor is able to discriminate properly some of the shapes, even when they could share some features. Such are the cases shown in the 4-th and 11-th rows that share a triangle structure at the bottom, and yet, they retrieve relevant elements in the first positions. However, a few shapes resulted more difficult to be properly ranked, specially those having very complex visual structures, e.g., the queries in the 5-th and 8-th rows in Figure 7.5.

Note that by considering as valid queries those instances with only few relevant documents, might generate results where the relevant documents are retrieved away from the beginning of the ranking list. Such is the case of the 3-rd row in Figure 7.5, where non-relevant documents are ranked first than relevant instances. However, they share visual similarity in this case.

In general term, the results show that the HOOSC descriptor has the potential to handle shapes from other archaeological sources.

## 7.3 MPEG-7 dataset

The second dataset we used in this chapter corresponds to the MPEG-7 Core Experiment CE-Shape-1 test set [Latecki et al., 2000], which is publicly available[2]. This dataset is a well known benchmark to compare the performance of shape descriptors as it is widely used for

[2] http://www.dabi.temple.edu/~shape/MPEG7/

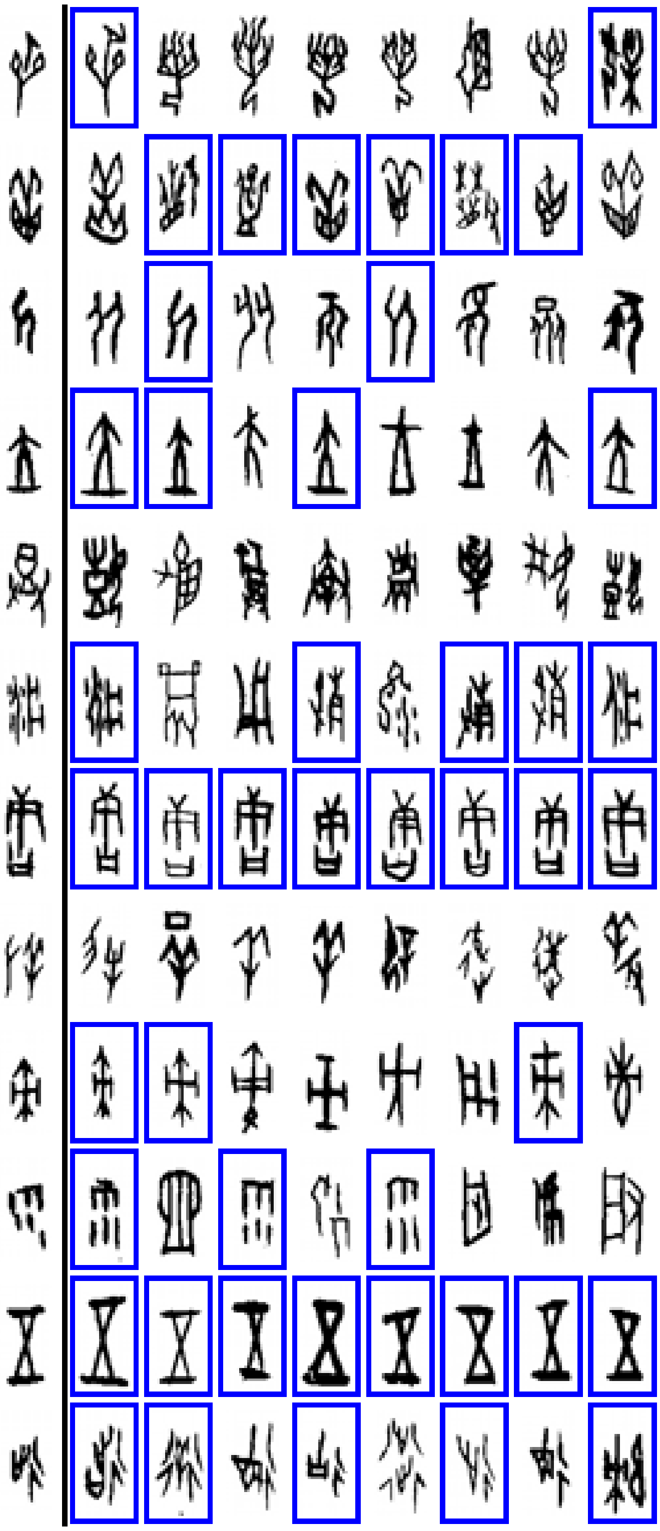


Figure 7.5: Example of retrieval results obtained with the HOOSC descriptor. The first column shows 12 queries randomly selected, and the remaining columns the 8 most similar instances according to the HOOSC descriptor. Relevant instances are highlighted with a blue rectangle.

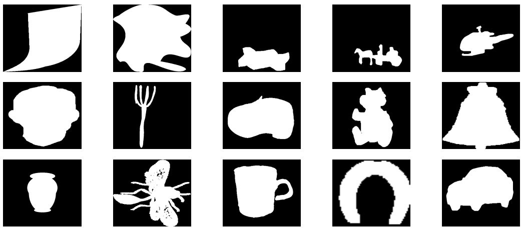


Table 7.2: 15 examples of the MPEG-7 dataset.

retrieval experiments of shape images, and the performance of several descriptors is reported on a website. Table 7.2 shows visual examples of the instances in this dataset.

Different from the complex shapes of the Maya and Chinese instances, the MPEG-7 contains relatively simpler shapes. More specifically, it contains silhouettes mainly composed of components having single- closed-contours. This dataset contains instances of generic objects, such as planes, animals, cars, insects, mugs, etc. Examples of the MPEG-7 dataset are shown in Figure 7.2. In total, it comprises 70 classes of silhouettes with 20 instances each. We divided it randomly at the same rate as the syllabic Maya dataset: 80% $G_C$ and 20% $G_Q$.

### 7.3.1 Experimental settings

The original HOOSC descriptor was not designed to handle rotation nor reflexion. Thus, we started by manually aligning 283 instances (≈20%) that were either rotated or reflected. We refer to this version of the data as A-MPEG-7.

Experimentally, we noticed that the HOOSC variants tested with the Maya glyphs did not achieved as good results with the A-MPEG-7 dataset, this is due to the different visual nature of the instances. Therefore, we tested the retrieval performance of the HOOSC variants summarized in Table 7.3 and compared them with the Generalized Shape Context (GSC) on the (aligned) A-MPEG-7 dataset. Note that the variants HOOSC0 and HOOSC4 in Table 7.3 correspond to the same variants shown in Table 5.2 of section 5.5.

Later, we tested the HOOSC variant that achieved the best retrieval performance on the original unaligned MPEG-7 dataset. Furthermore, in order to handle the rotation issue, we followed the same approach used in [Belongie et al., 2002] where the tangent vector at each pivot is treated as the positive $x$-axis of the reference polar grid, thus resulting in a rotation-invariant descriptor.

Previous works on retrieval of shapes tested on this dataset are compared via the *Bull's eye* score (*bes*) [Bai et al., 2010], which is the ratio of the total number of shapes from the same class retrieved among the 40 most similar positions to the highest possible number (i.e., 20 in the case of the MPEG-7 dataset). Thus, this metric is equivalent to the recall at top 40 retrieved elements. We present our results in terms of both *mAP* and *bes*. The use of *mAP* makes the

| HOOSC | 0 | 5 | 6 | 7 | 4 |
|---|---|---|---|---|---|
| Points from | Raw | Raw | Raw | Raw | Thin |
| Pivots w.r.t. | Pivots | Points | Points | Points | Points |
| Rings | 1:5 | 1:5 | 1:2 | 1:5 | 1:2 |
| self-position | NO | NO | NO | YES | YES |

Table 7.3: HOOSC variants evaluated in the A-MPEG-7 dataset. The first row indicates a version of the HOOSC descriptor. We use that number to refer to the versions as HOOSCX (e.g., HOOSC5). Improvements are highlighted in blue. Note that when only *rings* 1 and 2 are used, this implies a reformulation of the distance intervals of the polar grid (see chapter 5.2).

| Classes | GSC | HOOSC | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 4 | 5 | 6 | 7 |
| *mAP* | 0.813 | 0.848 | 0.790 | 0.852 | 0.849 | **0.867** |
| *bes* | 0.882 | 0.905 | 0.848 | 0.908 | 0.906 | **0.918** |

Table 7.4: *mAP* and *bes* for the methods evaluated in the A-MPEG-7 dataset.

results comparable with those obtained in chapter 5.

### 7.3.2 Retrieval results

Table 7.4 shows the results for the GSC and the evaluated variants of the HOOSC on the aligned A-MPEG-7 dataset.

The HOOSC4 which works the best with the Maya hieroglyphs does not perform as well with the A-MPEG-7 dataset, as shown in Table 7.4. The reason is due to two factors: a) as these shapes are dominantly filled and clean convex silhouettes with very well defined boundaries (rather than shapes with complex details), the morphological thinning transformation results in a loss of information and in descriptors with lower discriminative power. Also, sampling the pivots directly from the contours produces better results for these shapes, as shown with the HOOSC5 to HOOSC7 in Table 7.4; b) unlike the Maya hieroglyphs, where using the 5 rings (the whole spatial scope) adds noise to the description, using the 5 rings with the A-MPEG-7 silhouettes does not harm the description and the retrieval performance remains competitive as shown by HOOSC5 and HOOSC6. We can see that computing descriptors for pivots with respect to the whole set of points, and incorporating the relative self-position in the descriptor provide good results (HOOSC7).

Finally, we incorporated robustness to rotation, and experimented with the original unaligned shape instances of the MPEG-7 dataset, achieving results of 0.733 of *mAP* and 0.811 of *bes* with the HOOSC7. Examples of this experiment are shown in Table 7.5, we can see that while rotated instances are well retrieved, the HOOSC is not robust to reflected shapes yet.

As already mentioned, the results of previous methods evaluated in this dataset are public and reported in terms of the *Bull's eye* score (*bes*). Comparing the performance of previous

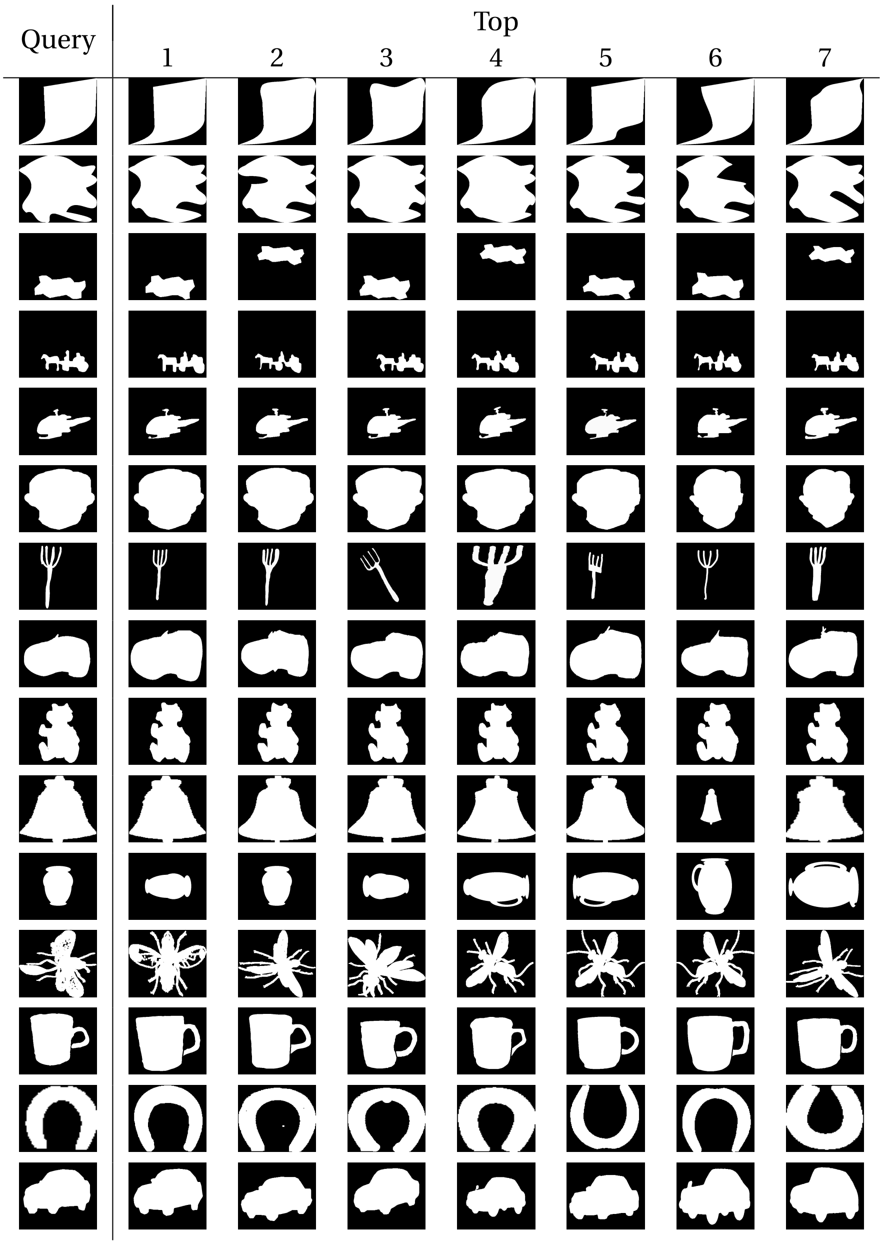

Query
Top
1
2
3
4
5
6
7

Table 7.5: 15 queries randomly chosen from the MPEG-7 dataset and their corresponding Top 7 retrieved candidates via the HOOSC7 method.

| Method | SC | HOOSC | [Bai et al., 2010] |
|---|---|---|---|
| *bes* | 0.765 | 0.811 | 0.933 |

Table 7.6: Comparison of retrieval results previously reported and the HOOSC descriptor for the MPEG-7 dataset.

methods with the performance achieved using the HOOSC7 in terms of *bes*, we can see that its performance is competitive with those methods. More precisely, the performance of HOOSC ranks among the top 12 reported methods. Table 7.6 shows the *bes* results reported on the MPEG-7 dataset[3] for the SC descriptor and the method with the best results [Bai et al., 2010], which applies a local diffusion process to improve the similarity measure of each pair of shapes.

Recalling section 5.2, it is important to mention that the quality of the hieroglyphs varies drastically due to the nature of the documents from which they are extracted. Thus, some improvements of the HOOSC descriptor (thinning and using only rings 1, 2) are specifically designed to deal with noise. In contrast, the MPEG-7 shape dataset is cleaner.

## 7.4 Conclusions

In this chapter, we validated the generalization of the HOOSC descriptor evaluating it on two dataset of shapes taken from sources different to the Maya hieroglyphs; the Oracle-Bones Inscriptions (OBI), which is another example of cultural heritage visual material, and the MPEG-7 Core Experiment CE-Shape-1 test set (MPEG-7).

Our results on the OBI dataset show that the use of statistics about local orientation information is of high importance to achieve accurate shape description. This is demonstrated as the HOOSC descriptor achieved better retrieval results than both the SC, which does not incorporate local orientations, and than the GSC, which only uses dominant orientations.

We also found that three out of the five features of the HOOSC descriptor are suitable to describe the convex and clean silhouettes of the MPEG-7 dataset. More specifically, the description of pivots with respect to points, the use of the histogram of local orientations, and the incorporation of the relative position of the pivot, are all features that help improve the retrieval performance on this dataset. In contrast, the use of thinning algorithms for convex silhouettes and the constraining of the spatial scope of the description result on loss of relevant information.

Although the HOOSC was not designed to handle rotations, we found that such a feature is easy to incorporate, as shown in the results with both datasets. Future work is still required to provide the HOOSC descriptor with robustness against reflection.

[3] http://www.dabi.temple.edu/~shape/MPEG7/

Finally, as a first step towards an application to archaeology, we implemented a version of a retrieval machine that we expect to become useful for the work of archaeologists and epigraphists. When such a tool is improved, we hope it will ameliorate the amount of time invested by archaeologists in manual searches, and it will help in the training of scholars who learn about the Maya writing system.

# 8 Towards automatic detection of Maya Syllables

As we mentioned in chapter 2, the translation and interpretation of Maya inscriptions require the identification of their components (syllabographs, logographs, and narrative elements). Currently, this identification process is manually done by experts, which often need to consult catalogs or to compare the context with others inscription previously deciphered. Having a mechanism capable of automatic detection of the writing elements could diminish the time required for the decipherment.

The automatic detection of hieroglyphs in large inscription is a difficult task. First, because it requires methods to describe the symbol of interest and produce a representation that is consistent between the query and its instance potentially present in the inscription. These two representation might vary as the query could be a segmented (and cleaned) image, whereas its instance in the inscription usually would be surrounded by other elements that might add “noise” to the description. Second, because it requires a method to efficiently scan the inscription and to find the correct matches.

Approaches for image detection work well on gray-scale images. Usually, they rely on variants of the sliding window approach [Viola and Jones, 2001, Lampert et al., 2008], in which usually a binary classifier evaluates a large set of subwindows and decides whether or not the element of interest is present on them. Often, the images are described using local descriptors such as SIFT [Lowe, 2004] or HOG [Dalal and Triggs, 2005], or by template matching approaches [Viola and Jones, 2001], and supervised methods are used to train the classifier.

In this chapter we present our initial approach to detect segmented syllabic instances in larger inscriptions of Maya hieroglyphs, and presents the results obtained using SIFT and HOOSC descriptors combined with different interest points. The chapters is organized as follows. Section 8.1 introduces the points of interest we used. Section 8.2 briefly reviews the SIFT descriptor. Section 8.3 introduces the dataset used for detection experiments. Section 8.4 explains the ad-hoc protocol used to approximate detection experiments. Section 8.5 presents our results. Finally, in section 8.6 we present our conclusions and directions for future work.

## 8.1 Interest Points

Two of the most popular Interest Points detectors used for local description of intensity images are the Difference-of-Gaussians (DoG) that detects blob-like regions, and the Harris corner detector. We briefly describe them in this section.

### 8.1.1 Blob-like regions

The Laplacian of Gaussians approach detects blob structures in intensity images by looking for maximum responses in a 3-D scale-space [Lindeberg, 1998]. This 3-D space results after the convolution of the image of interest $I(x,y)$ with a 2-D Gaussian filter that helps removing noise,

$$L(x,y;\sigma) = g(x,y,\sigma) * I(x,y), \tag{8.1}$$

where $*$ correspond to the convolution operator, and $g(x,y,\sigma)$ defines the Gaussian kernel at certain scale $\sigma$ and centered at $x$ and $y$,

$$g(x,y,\sigma) = \frac{1}{2\pi\sigma} e^{-\frac{(x^2+y^2)}{2\sigma}}. \tag{8.2}$$

After the Gaussian smoothing, the Laplacian of Gaussian ($LoG$) is computed by a second convolution with a Laplacian operator that attains strong responses for blobs of size$\sqrt{2\sigma}$,

$$LoG \equiv \nabla^2 L\left(x,y;\sigma\right) = \sigma\left(\frac{\delta^2 L}{\delta x^2} + \frac{\delta^2 L}{\delta y^2}\right), \tag{8.3}$$

where $\frac{\delta^2 L}{\delta x^2}$ defines the second derivative in the $x$ (or $y$) direction, and the multiplication by $\sigma$ is a normalization factor that counteracts the decay in the Laplacian response when the smoothing is increased [Crowley and Parker, 1984, Lindeberg, 1998].

The Laplacian of Gaussian can be efficiently approximated by the use of a scale-space pyramid constructed with Difference of Gaussians (DoG), in which the image of interest is smoothed with Gaussian filters at different scales and then subsampled, finally the local extrema are found by the subtraction of each two consecutive smoothed images [Lowe, 2004],

$$DoG \equiv L(x,y;k\sigma) - L(x,y;\sigma), \tag{8.4}$$

where $k$ is a factor to control the increment rate of the scales, and $L(x,y;\sigma)$ is computed as in Equation (8.1).

Both the LoG and DoG operators simultaneously detect the position ($x$ and $y$) and the characteristic scale ($\sigma$) of a blob structure.

### 8.1.2 Corners

The Harris method detects corner structures based on the autocorrelation matrix $A$,

$$A = g(x, y, \sigma) * \begin{bmatrix} \left(\frac{\delta I}{\delta x}\right)^2 & \left(\frac{\delta I}{\delta x}\right)\left(\frac{\delta I}{\delta y}\right) \\ \left(\frac{\delta I}{\delta x}\right)\left(\frac{\delta I}{\delta y}\right) & \left(\frac{\delta I}{\delta y}\right)^2 \end{bmatrix}, \tag{8.5}$$

which has two large positive eigenvalues when its center correspond to a corner structure [Harris and Stephens, 1988]. In practice, the corner measure can be efficiently computed by the combination of the determinant and the trace of the autocorrelation matrix as,

$$C = det(A) - \alpha trace^2(A), \tag{8.6}$$

where $\alpha$ is a constant to control the linear combination.

The Harris-Laplace corner detector is a hybrid method that combines the Harris corner detector with the Laplacian approach for selection of characteristic scale [Mikolajczyk and Schmid, 2004]. This method detects structures that are invariant to scale, location, and affine transformations. It has shown improved results in terms of repeatability of the detected points, and in the task of image matching.

## 8.2 SIFT Descriptor

The Scale Invariant Feature Transform (SIFT) descriptor is the most popular image description [Lowe, 2004]. This approach is efficient as it implements a cascade filtering approach, in which less expensive operations are performed on several candidate locations at initial steps along with filtering tests, and only those locations that pass the tests are carried out to the next step.

The SIFT image descriptor has 4 major steps:

1. **Scale-space extrema detection:** It detects candidate interest point by the use of the difference-of-Gaussians.

2. **Keypoint refinement:** Discards the candidate interest points with low contrast by evaluating the second order Taylor expansion of the DoG, and keeping only those points above certain threshold. It also discards points with poorly defined peaks, which correspond to those points whose principal curvature across the edge is much larger than the principal curvature along it. Thus, they can be found by evaluating the ratio $r = trace^2(A)/det(A)$, where $A$ is the autocorrelation matrix defined in Equation (8.5).

3. **Local orientation assignment:** Each final interest point is assigned a local orientation and magnitude at its characteristic scale. Both the orientation and magnitude are computed from the Gaussian-smoothed image.

4. **Description:** Each interest point is finally described by the distribution of orientations in its neighborhood computed at its characteristic scale. Roughly speaking, the region around the interest point (up to its scale) is dived into a 4x4 grid, and for each cell a histogram of orientations is computed. To avoid hard binning and boundary effects, the contribution of the local orientations is weighted inversely proportional to its distance to the center of each cell. Thus the final descriptor is 128-dimensional. A final normalization to unit length is applied to the descriptor.

As the SIFT descriptor was designed to deal with changes in intensity, there is no guaranty that it will work well on binary images. We decided to evaluate its performance, and use it as baseline for future investigations.

## 8.3 Data: Block Segments

The dataset used for these experiments consists in blocks that were randomly segmented from the large inscriptions that the project AJIMAYA has collected (see chapter 2). We decided to use these random blocks to have a better control over the experimental setup, as the inscriptions are sparsely annotated with respect to their very large size.

More specifically, we manually chose 10 different instances for each of the 24 most popular syllabic classes of Maya hieroglyphs. We labeled them as *ground-truth* and annotated their location. Then we generated the random blocks by segmenting the subwindows containing the ground-truth, 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 ground-truth, while the top and bottom margins have random sizes between one and four times the height of the corresponding ground-truth. We decided to use such sizes to generate random blocks containing enough visual information around the ground-truth, such that the challenge of a realistic detection setup is kept. Figure 8.1 shows one inscription with a random block and its corresponding ground-truth. And Figure 8.2 shows the details of the random block.

Later, we generated 20 variants of each ground-truth by randomly shifting their bounding boxes up to 0.2 times its width and height, and we annotated the location of these variants. This resulted in 200 new instances per syllabic class that we labeled as *positive* instances. Finally, for each of the selected blocks, we annotated the location of all the remaining bounding boxes that do not overlap with the ground-truth, and that are of the same size as the ground-truth in its corresponding random bock. This procedure resulted in 6000+ bounding boxes that we labeled as *negative* instances. This set ranks from 0 to 40 negative instances per block, with an average number of 26.1 ± 16.0.

By annotating the images this way, we turned the traditional detection approach based on sliding windows into a retrieval-based approach. This change avoids the possibility of detecting true-non-annotated instances, which would impact the performance in our experiments. This annotation also resulted in fast detection experiments, although at the price of non-exhaustive

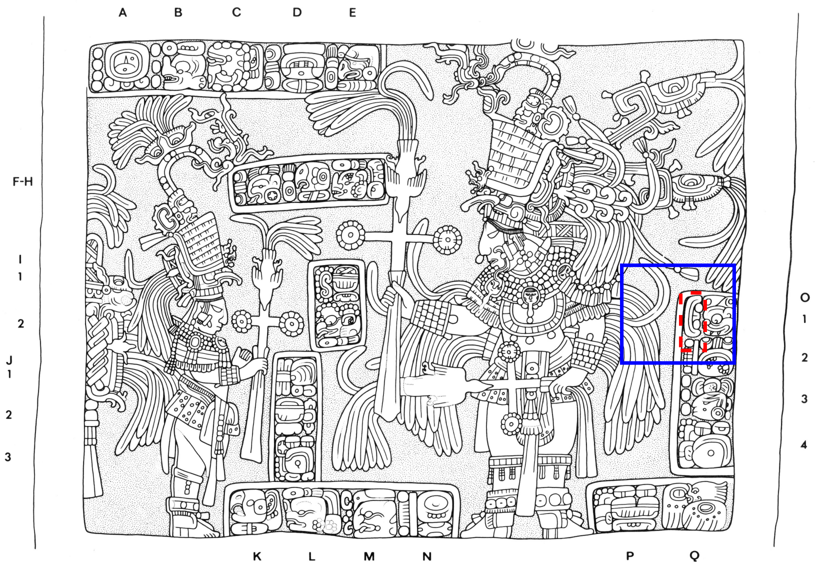


Figure 8.1: Maya inscription found in a lintel in Yaxchilan, and containing syllabographs, logographs, and iconography. Highlighted in a dashed red square is the ground-truth corresponding to the random block delimited by the blue rectangle.

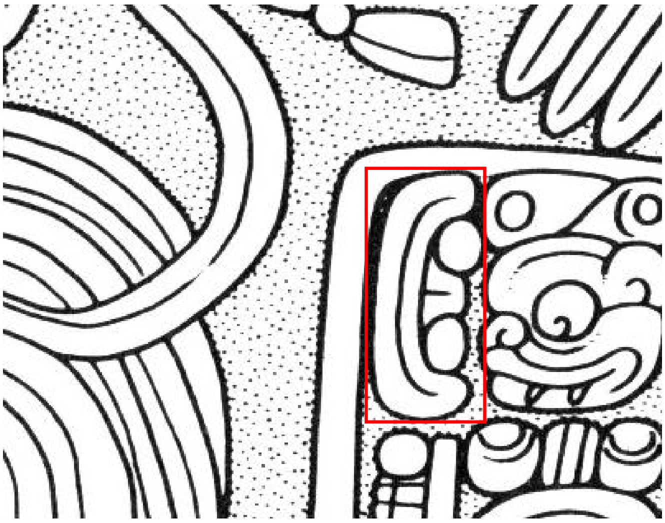


Figure 8.2: Random block extracted from the inscription shown in Figure 8.1. The ground-truth is highlighted in red and corresponds to syllable *u*.

| 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 |

Table 8.1: Tested combinations of interest points and local descriptors for detection of Maya syllables.

scanning of the images.

In summary the dataset contains: 240 *ground-truth* instances (10 for each of the 24 most popular syllabic classes), 4800 *positive* instances (200 for each of the 24 most popular syllabic classes), and 6000+ *negative* instances that do not belong to any of the positive classes (the negative instances might contain logograms, noise, or any other kind of Maya inscriptions).

## 8.4 Experimental setup

We conducted detection experiments using DoG and Harris-Laplace interest point, and SIFT and HOOSC descriptors. Table 8.1 summarizes the combinations we evaluated.

The HOOSC used for detection corresponds to a version that includes three of the five characteristics of the descriptor. Namely, it uses: (1) thinned contours, (2) description of pivots with respect to points, (3) the histogram of local orientations in each region. Since the size of the visual elements in the random blocks are unknown, it is not possible to compute the spatial scope of the descriptor as a function of the pairwise distances of the contour points. Therefore, we made use of the characteristic scale of the interest point at which the descriptor is computed. More precisely, we used the two most local rings with boundaries at 0.5 and 1.0 times the characteristic scale of the interest point. We did not use the explicit relative position of the point in the description because of the issue of the unknown size, and also because some elements might be rotated within the inscriptions.

As shown in Table 8.1, we decided to compute SIFT descriptors of thinned versions of the shapes, this is done with purposes of comparison since HOOSC uses shapes with thinned contours as explained in chapter 5. The DoG and SIFT implementations correspond to the libraries provided in OpenCV, whereas the Harris-Laplace was implemented in Matlab. Namely, we performed our experiments under the following five-steps protocol:

1. **Interest point detection:** The first step of our methodology consisted in detecting points of interest (DoG or Harris-Laplace), as well as their characteristic scales and local orientations in the blocks randomly selected, i.e., we did not consider individually each of the annotated bounding boxes but the block as a whole, thus avoiding boundary

effects that are not expected in a common detection setup.

2. **Description:** Then, we computed the corresponding local descriptors (SIFT or HOOSC) using their characteristic scales and local orientations.

3. **Indexing:** After computing the descriptors for the blocks, we estimated visual vocabularies by using the $k$-means clustering algorithm, and then we constructed *bov* representations for each of the bounding boxes. The *bov* were constructed taking into account only those points whose scale is 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, e.g., points whose scale is much larger than the bounding box, and points close to the edge whose scale intersects only a small proportion of the bounding box. More specifically, we exclude all those points whose intersection ratio $r = A/(2s)^2$ was below 0.5, where $s$ is the characteristic scale of a given point, and where $A$ is the intersection area between the characteristic scale and the current bounding box.

4. **Detection:** Once the bounding boxes are described, we computed the euclidean distance from the *bov* of each ground-truth against the *bov*s of all the positive and negative bounding boxes belonging to the blocks of the same class of 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 block. Note that we excluded the block that contains the current ground-truth, as itself and all its positive variants are easily detected. Namely, each ground-truth is expected to have smaller distances to the 189 (ground-truth + positive) bounding boxes than to the other negative instances (234.8 negative bounding boxes on average). Thus our detection method is not a classical exhaustive sliding window but an approximation based on a retrieval approach.

5. **Evaluation:** Finally, we ranked all the bounding boxes based on the computed distances. We present the detection performance based on

   - 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. Where the detection rate $DR$ is compute as the fraction of true positives $TP$ to the total number of positives,

     $$DR = \frac{TP}{TP + FN}, \tag{8.7}$$

     where, $FN$ denotes the set of false negatives. And the false-positive-rate-per-window $FPPW$ is the ratio of false positives $FP$ to the total number of negatives,

     $$FPPW = \frac{FP}{FP + TN}, \tag{8.8}$$

     where, $TN$ denotes the set of true negatives.

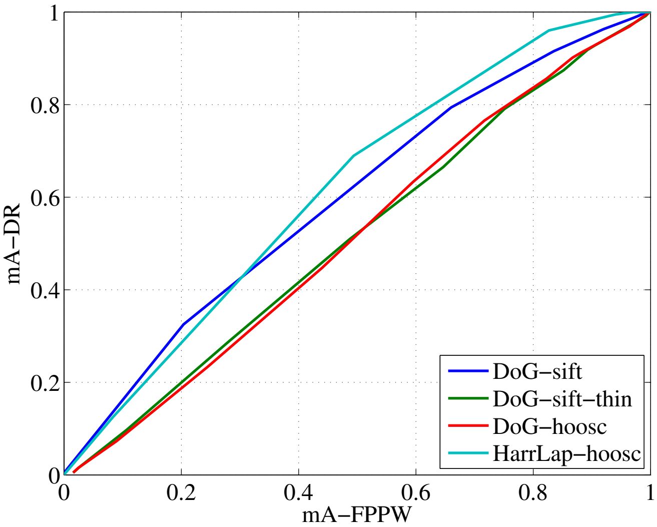


Figure 8.3: ROC curves.

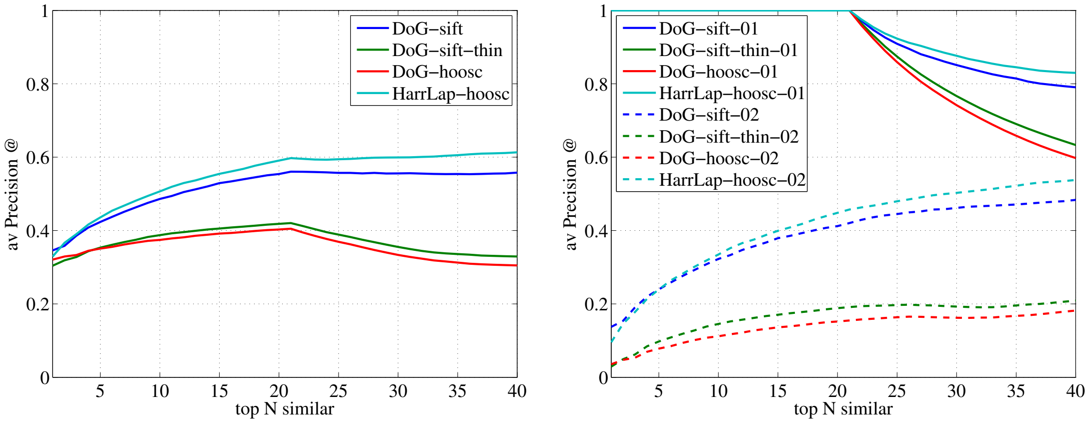
- Curves showing the average-precision achieved at different top-N positions of the ranked subwindows.
- The mean Average Precision *mAP*.

## 8.5 Results

The ROC curves in Figure 8.3 show that the use of DoG points with thinned contours gives detection rates close to chance, both with SIFT and HOOSC descriptors. This observation is not especially surprising as binary images lack of intensity information which is the main clue to estimate DoG interest points. The motivation to attempt the use of DoG points in thinned shapes was based on the high frequency of blob structures present in the Maya syllables. In practice, DoG points also detected large blob structures that encompass visual information beyond the locality of the glyph of interest. The exclusion of such large structures to avoid adding noise to the glyph's description, as explained in section 8.4, resulted in inaccurate shape representations. Note that the detection rate is relatively increased with the estimation of the DoG points on the original shapes with thick contours (blue curve in Figure 8.3), this is mainly explained by the used of the Gaussian convolutions that smooth the thick contours and approximate intensity changes.

The use of Harris-Laplace interest points resulted in a slightly increased detection rate when used on thinned structures, as shown in the HarrLapl-HOOSC curve on Figure 8.3.

The relative difference among the four methods in terms of average precision remains proportional to their differences in the ROC curves, as shown in 8.4a. The slight peak in the retrieval precision at position 21 results because some classes have very similar instances, such that for a given ground-truth that is queried, the 21 bounding boxes (ground-truth + positives) of (at



(a) Average precision at top N

(b) Split average precision at top N

Figure 8.4: (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.

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 groups: one with the curves whose precision remains equal to 1 at the 21-st position, and the other with the remaining vectors. The solid curves in Figure 8.4b (named XXX-01) show the average precision for the first sets, whereas the dashed curves (named XXX-02) correspond to the average of the second sets. An examples of this case is also illustrated in the first row of results in Figure 8.5, where the query (in blue) is well detected at slightly shifted locations in one its relevant random blocks. In practice, 18 of the 21-st most similar bounding boxes for this query were detected in the same random block, and the other 3 in another relevant random block. Furthermore, this is the query with the highest precision.

The cyan solid curve in Figure 8.4b 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 shape descriptor works well for certain Maya syllabic classes. Figure 8.5 shows detection results obtained with the this method.

Finally, Table 8.2 summarize the performance of the four tested methods in terms of mean average precision.

Overall, the use of corners as interest points for describing shapes seems to achieve better performance over blob structures.

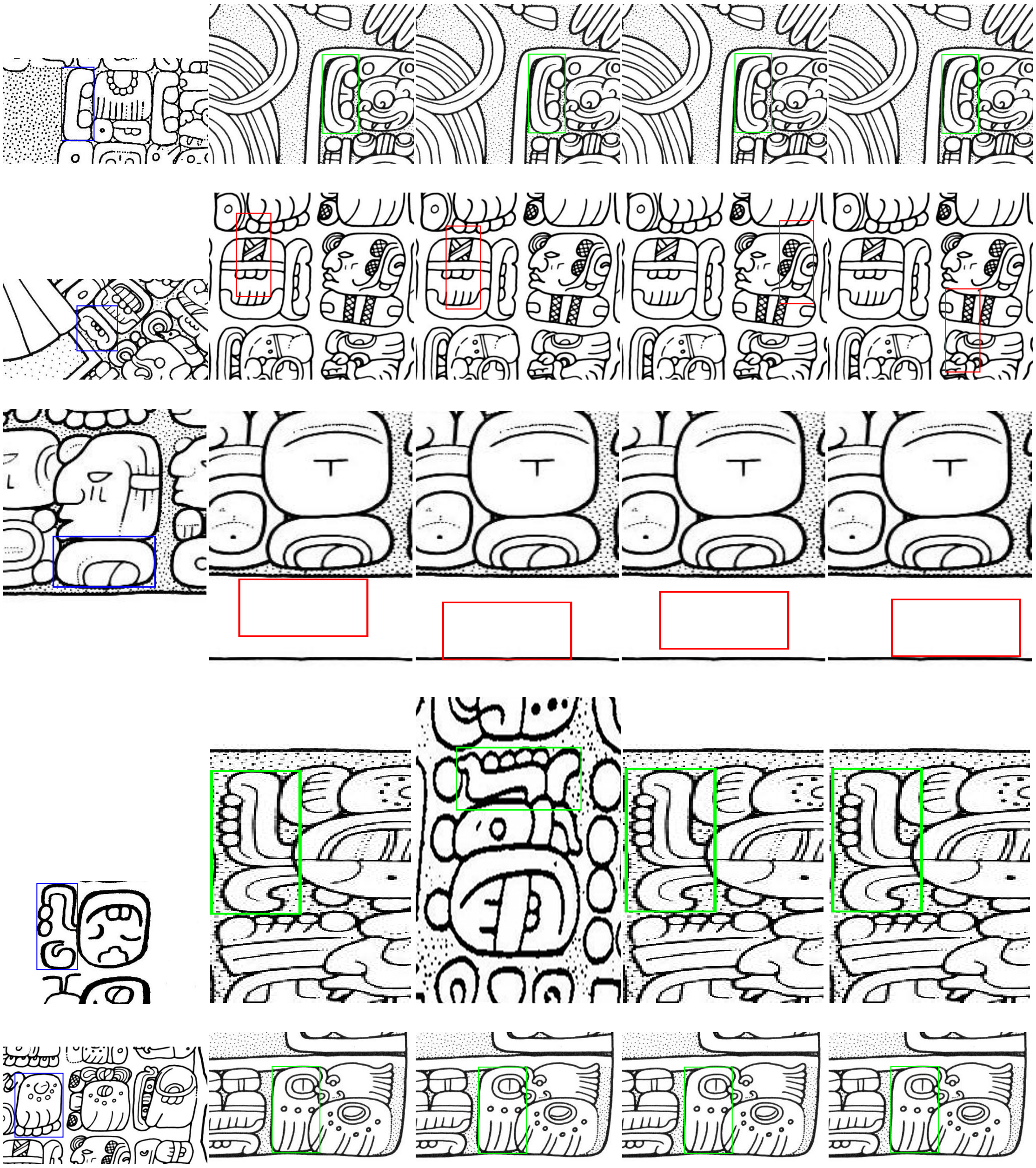


Figure 8.5: Visual examples of detection with Harris-Laplace interest points and HOOSC descriptors. The first random block in each row contains a query inside a blue rectangle (ground-truth). The next four random block correspond to the four most similar bounding boxes according to the method, where green rectangles indicate correct detection, and red rectangles indicate erroneous detection.

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

Table 8.2: Mean average precision for the combinations of interest points and local descriptors tested for detection of Maya syllables.

## 8.6 Conclusions

In this chapter we explored an initial approach towards detection of syllabic Maya hieroglyphs in larger inscriptions. We evaluated DoG and Harris-Laplace interest points combined with SIFT and HOOSC descriptors for this task.

To avoid heavy the computations of an exhaustive scanning of large inscriptions, and the detections of positive instances that have not been manually annotated yet (which requires expert knowledge), we relied on a controlled setup that uses synthetic data extracted from the large inscriptions. More specifically, we generated positive instances by annotating the ground-truth locations of certain syllables and the locations of slight shifts of their bounding boxes. We also generated negative instances by annotating the location of bounding boxes without overlap with the positive instances. Then, we performed comparison of *bov* representations of the bounding boxes, and approximated a sliding windows detection approach based on retrieval based experiments.

Our results show that regardless of the local descriptor, the use of DoG points with thinned contours gives detection rates close to chance, which results as a consequence of the lack of intensity information in binary images. A slightly better performance is achieved by using ticker contours as the Gaussian convolutions smooth, thus generating images with some sort of intensity changes. In practice, the use of corner detectors seems suitable for local description of binary images, as the best results were obtained by using the Harris-Laplace interest points.

In this initial exploration we evaluated to interest point detectors that are designed for gray-scale images. The results suggest the need for interest point detectors tailored for binary images. Although the current experimental setup was useful in this initial stage, it did not achieved very good result for some of the visual classes, indicating that the direct comparison of *bov* representations is not the best choice for image detection. Also, this detection approach would turn unnecessary once enough data is collected, as gathering more data could allow for using supervised learning approaches that increase the performance in the detection of these complex shapes. However, note that laborious manual annotation is require in order to collect enough data, as fully annotating one inscription requires the manual identification and labeling of all the inscribed syllabographs and logographs, as well as all the empty areas of the image.

The detection rate could also be improved by further exploration of other types of interest points, as well as more suitable approaches to estimate characteristic scales for points in shapes.

# 9 Conclusions

In this thesis we have proposed and investigated Computer Vision techniques for Content-Based Image Retrieval (CBIR) of complex shapes. More specifically, we have contributed to the state-of-the-art in the field of statistical shape description of complex shapes, validating our methods on datasets of images depicting shapes with historical value. In this chapter we summarize our research and contributions, and we also discuss open issues and potential future directions of our research.

## 9.1 Contributions

In **Chapter 2**, we described the visual complexity of the Maya writing system, and why it poses non trivial challenges to the tasks of statistical description and automatic identification. We also explained the manual process required to generate digital datasets of Maya inscriptions, and the urgent need to develop computational tools that could support some steps of such a laborious process, thus helping archaeologists in their daily work.

In **Chapter 3**, we discussed related work in the areas of CBIR, shape description and retrieval, indexing techniques such as quantization and sparse coding, the detection of points of interest for image description and estimation of characteristic scales, and shape-based image detection. We also reviewed previous works that have applied Computer Vision techniques in the fields of cultural heritage and digital art preservation.

In **Chapter 4**, we reviewed the Shape Context (SC) descriptor. We evaluated its performance in retrieval experiments on an initial dataset of Maya hieroglyphs that ranks in the order of hundreds. We proposed the use of an adaptive sampling method to select the subset of points used for shape description, allowing for better covering of the shape structure and more accurate description when compared with the fixed size of the sampling process that SC traditionally uses. We improved the cost function and the dissimilarity index that Shape Context uses to retrieve and rank shapes, and showed that the use of dummy handlers is important for shape matching when two shapes differ in size or resolution, and specially when they share similar overall structures but have different internal details. Our proposed methodologies resulted in the improvement of the retrieval precision with respect to the use of the traditional Shape Context approach.

In **Chapter 5**, we introduced the Histogram-of-Orientations Shape-Context (HOOSC), a new shape descriptor designed to overcome some of the limitations of SC. We evaluated its descriptive power with retrieval experiments performed on a dataset of 1200+ Maya syllabic hieroglyphs. Our results show that the adequate combination of the HOOSC features allows for improved retrieval results over SC and its variants. More specifically, HOOSC achieved a 18.8% absolute improvement in terms of retrieval precision over existing methods. We also validated the potential of HOOSC as tool for archaeological analysis, using it to statistically analyze the visual variations of Maya hieroglyphs. We found that Maya hieroglyphs tend to have less visual variability in subsequent periods, suggesting a gradual convergence, and that their visual variability increased in regions towards the borders of the ancient Maya territory, suggesting that these symbols got enriched with new visual features when they disseminated.

In **Chapter 6**, we investigated two different quantization approaches to construct efficient bag-of-words representations (*bov*). Namely, we compared the *k*-means clustering algorithm with the K-SVD sparse decomposition technique. We found that depending on the given application, sparse techniques might not outperform simpler clustering algorithms. We also confirmed previous observations regarding the use of max-pooling as the strategy that provides the best results. We proposed a thresholding method that facilitates the sparse decomposition of signal with certain amount of noise, showing that by using this approach the retrieval performance can be improved.

In **Chapter 7**, we present an evaluation of the HOOSC descriptor that validates its potential to be used with shape images of different nature to the Maya syllables. Namely, we performed retrieval experiments on the MPEG-7 Core Experiment CE-Shape-1 test set, and on a non public dataset of ancient Chinese characters (OBI). The results on the MPEG-7 dataset confirmed that the use of the distinct features of the HOOSC descriptors are dependent on the dataset, as some of them might not be required when the shape instances lack internal details or have convex contours as main global feature. On the OBI dataset, the HOOSC achieved better retrieval performance for most of the classes when compared with other shape descriptors, and it performed best on average. This chapter also introduces the first version of a system developed to retrieve Maya hieroglyph.

In **Chapter 8**, we presented an initial approach we followed to perform detection of segmented syllabic glyphs in larger inscriptions. This approach is an ad-hoc methodology that allows the approximation of detection experiments given the current limitations in terms of available data. The chapter also describes the interest point detectors (difference-of-Gaussians and Harris-Laplace) and shape descriptors (SIFT and HOOSC) that we used. Our preliminary investigation has shown not difference among the many tested combinations, and suggests the need for a more adequate approach for the task of shape detection in binary images.

## 9.2 Open issues and future work

Besides the contributions presented in this dissertation, several other computer vision tasks remain to be investigated, both for shape analysis and in concrete applications to cultural heritage. In this section, we comment open issues that we identified during our research, such as the potential improvements in shape description that supervised learning could provide for retrieval and classification of Maya hieroglyphs, which in turn could be also used for detection of hieroglyphs. Also related to shape-based detection, there is the problem of estimating characteristic scales for shapes descriptors. We also comment the future lines of research that could be explored for other components of the Maya writing system and for larger datasets.

The visual features that HOOSC recovers are consistent among instances of the same visual classes. We believe that using **supervised learning**, this features could lead to more structured descriptions, which in turn could improve the retrieval and detection performance, and that could be used with purposes of shape-based image classification.

With no doubt, further investigation in **detection** of shapes is needed given the current results. The manual detection of hieroglyph requires expert knowledge and can be quite time consuming for large inscriptions, thus the automatic or semi-automatic detection of shapes is a problem of special interest for archaeologists. Related to the task of detection, there is the need to automatically estimate **characteristic scales** for SC and HOOSC descriptors. The current formulations of these shape descriptors define their spatial context (scale) as a function of the distance among the input set of points, which works well for segmented instances whose size is know a priori. However, this is no longer true for detection applications, where the size of the potential instances in the inscription is unknown. Although we explored techniques to both detection and scale estimation, further investigation of them is still required.

On a different direction, the design and collection of larger dataset is one of the principal tasks to be performed for future work. In the most general case, having more instances of each of the visual classes will open the door to use more sophisticated techniques based on supervised learning that could improve both the retrieval and detection of Maya hieroglyphs, and that in turn could also be the basis to investigate the classification of these complex shapes. In a more concrete setup, having larger numbers of instances for each combination of visual class - temporal window - Maya region is a necessity in order to investigate in depth the visual variations and evolution of the Maya writing system.

Throughout our work, we investigated the shape description of segmented Maya syllables, leaving aside the logographs that represent about 80% of the known Maya corpus, and that might contain even more complex structures or that might share higher portions of visual patterns, thus requiring more refined methods for accurate description. In a related direction, the use of glyph-blocks as basic units of analysis is of great interest for the archaeological community, and will pose new challenges as we would have to deal with common issues of the Maya writing system such as conflation, infixation, superimposition, and pars pro toto. Using glyph-blocks will also likely require the incorporation of new techniques such as Markov models to address the co-occurrences phenomenon. Furthermore, the discovery of visual patterns in glyph-blocks could provide with statistical information to address the problems of detection and segmentations of single glyphs and new glyph-blocks.

Finally, our research was conducted on datasets of binary images, as this type of data corresponds to the format most commonly used by archaeologists in their own research, and we have targeted the design of tools that could help them do so. However, the amount of data in other formats containing color and texture information is also vast, and requires the design of tools to deal with it. Examples of this type of data are photographies of vessels and of ancient "paper" (codices), whose analysis could contribute to historical and archaeological research with insights of different nature to those concluded after analyzing stone inscriptions.

## 9.3 Final thoughts

Overall, this dissertation represents an initial step towards the systematic integration of Computer Vision techniques in the analysis of ancient Maya cultural heritage. Our interdisciplinary approach is at once challenging and rich, as it genuinely addresses needs and open problems in archaeology and epigraphy, and poses a number of questions for computer vision research, resulting in a unique opportunity to integrate knowledge in computing, archaeology, and epigraphy. According to feedback received by our archaeology partners, the implementation of this kind of tools will be helpful to study the evolution of the Maya writing system, to analyze visual relationships among symbols, and to categorize hieroglyphs recently discovered. We believe that further investigation under this multidisciplinary format will result in systems to support the work of scholars in archaeology, and to disseminate information to general audiences like visitors to museums and online catalogs.

Finally, our work has been the starting point for a larger and more ambitious project, which will be conducted as a collaboration between the Idiap Research Institute, the University of Geneva, and the University of Bonn, starting from early 2013, in which many of the ideas discussed here will be explored.

# Bibliography

Foundation for the Advancement of Mesoamerican Studies, Inc. URL http://www.famsi.org.

H. Aanæs, A. L. Dahl, and K. S. Pedersen. Interesting Interest Points. *International Journal of Computer Vision*, 97(1):18–35, March 2012.

S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *IEEE International Conference on Computer Vision*, 2009.

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, November 2006.

R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., 1999.

A. D. Bagdanov, A. Del Bimbo, L. Landucci, and F. Pernici. MNEMOSYNE: enhancing the museum experience through interactive media and visual profiling. In *Proceedings of the First International Workshop in Multimedia for Cultural Heritage*, 2011.

X. Bai and L. J. Latecki. Path Similarity Skeleton Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1282–1292, July 2008.

X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu. Learning Context-Sensitive Shape Similarity by Graph Transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5): 861–874, 2010.

F. Banfi and R. Ingold. Computing Dissimilarity Between Hand-drawn Sketches and Digitized Images. In *Proceedings of the Third International Conference on Visual Information and Information Systems*, June 1999.

S. Belongie, J. Malik, and J. Puzicha. Shape Context: A new Descriptor for shape matching and object recognition. In *Neural Information Processing Systems Conference*, December 2000.

S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, June 2002.

N. Boujemaa, V. Gouet, and M. Ferecatua. Approximate search vs. precise search by visual content in cultural heritage image databases. In *4-th International Workshop on Multimedia Information Retrieval (MIR'02) in conjunction with ACM-Multimedia*, December 2002.

Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2010a.

Y.-L. Boureau, J. Ponce, and Y. LeCun. A Theoretical Analysis of Feature Pooling in Visual Recognition. In *International Conference on Machine Learning*, June 2010b.

J. K. Browder. *Place of the high painted walls: The Tepantitla murals and the Teotihuacan writing system*. PhD thesis, University of California, 2005.

J Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, June 1986.

Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-Bag-of-Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2011.

J. L. Crowley and A. C. Parker. A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 6(2):156–170, February 1984.

N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

R. Datta, D. Joshi, J. Li, J., and Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 39(2):1–60, 2008.

A. Del Bimbo and P. Pala. Visual Image Retrieval by Elastic Matching of User Sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):121–132, February 1997.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Societey*, 39(1):1–38, 1977.

W. Duan, F. Kuester, J.-L. Gaudiot, and O. Hammami. Automatic Object and Image Alignment using Fourier Descriptors. *Image and Vision Computing*, 26(9):1196–1206, September 2008.

M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa. Sketch-Based Shape Retrieval. *ACM Transactions on Graphics*, 31(4):1–10, 2012.

J. H. Elder and S. W. Zucker. Local Scale Control for Edge Detection and Blur Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):699–716, July 1998.

V. Ferrari, F. Jurie, and C. Schmid. Accurate Object Detection with Deformable Shape Models Learnt from Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007.

V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of Adjacent Contours for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):36–51, 2008.

R. Flad. Divination and Power: A Multi-regional View of the Development of Oracle Bone Divination in Early China. *Current Anthropology*, 49(3):403–437, 2008.

Y. Frauel, O. Quesada, and E. Bribiesca. Detection of a Polymorphic Mesoamerican Symbol Using a Rule-based Approach. *Pattern Recognition*, 39:1380–1390, 2006.

A. Frome, Y. Singer, F. Sha, and J. Malik. Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification. In *IEEE International Conference on Computer Vision*, October 2007.

S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely - Laplacian sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

D. Gatica-Perez, E. Roman-Rangel, J.-M. Odobez, and C. Pallan. New world, New Worlds: Visual Analysis of Pre-Columbian Pictorial Collections. In *Proceedings of the First International Workshop in Multimedia for Cultural Heritage*, April 2011.

I. Graham. *Introduction to the Corpus, Corpus of Maya Hieroglyphic Inscriptions*. Peabody Museum of Archaeology and Ethnology, Cambridge, 1975.

N. K. Grube. *Die Entwicklung der Mayaschrift: Grundlagen zur Erforschung des Wandels der Mayaschrift von der Protoklassik bus zur spanischen Eroberung*. PhD thesis, Universität Hamburg, 1989.

C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference*, August 1988.

G. Heitz, G. Elidan, B. Packer, and D. Koller. Shape-Based Object Localization for Descriptive Classification. *International Journal of Computer Vision*, 84(1):40–62, August 2009.

M.-K. Hu. Visual Pattern Recognition by Moment Invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, February 1962.

J. M. Hughes, D. J. Graham, and Daniel N. Rockmore. Quantification of Artistic Style through Sparse Coding Analysis in the Drawings of Pieter Bruegel the Elder. *Proceedings of the National Academy of Sciences*, 107(4):1279–1283, January 2010.

T. Jiang, F. Jurie, and C. Schmid. Learning Shape Prior Models for Object Matching. In *Computer Vision and Pattern Recognition*, August 2009.

C. R. Johnson, E. Hendriks, I. Berezhnoy, E. Brevdo, S. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang. Image Processing for Artist Identification - Computerized Analysis of Vincent van Gogh's Painting Brushstrokes. *IEEE Signal Processing Magazine, Special Issue on Visual Cultural Heritage*, 25(4):37–48, July 2008.

F. Jurie and C. Schmid. Scale-invariant shape features for recognition of object categories. In *IEEE Conference on Computer vision and Pattern Recognition*, June 2004.

J. Justeson, W. Norman, L. Campbell, and T. Kaufman. *The Foreign impact on Lowland Maya Language and Script.* Middle American Research Institute Publication 53. Tulane University, New Orleans, 1985.

H. W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

Alfonso Lacadena. *Evolución formal de las grafías escriturarias Mayas: implicaciones históricas y culturales.* PhD thesis, Universidad Complutense de Madrid., 1995.

L. Lam, S.-W. Lee, and C. Y. Suen. Thinning Methodologies-A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):869–885, September 1992.

C. Lampert, M. Blaschko, and T. Hofmann. Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

F. Larue, M. Dellepiane, H. Hamer, and R. Scopigno. Automatic Texturing without Illumination Artifacts from In-Hand Scanning Data Flow. In *Proceedings of the First International Workshop in Multimedia for Cultural Heritage*, 2011.

F. Larue, M. D. Benedetto, M. Dellepiane, and R. Scopigno. From the Digitization of Cultural Artifacts to the Web Publishing of Digital 3D Collections: an Automatic Pipeline for Knowledge Sharing. *International Journal of Computer Vision*, 7(2):132–144, April 2012.

L. J. Latecki, R. Lakamper, and T. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2006.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, December 2007.

Y.J. Lee and K. Grauman. Shape Discovery from Unlabeled Image Collections. In *Computer Vision and Pattern Recognition*, August 2009.

P. H. Lewis, K. Martinez, F. S. Abas, M. F.A. Fauzi, S. C.Y. Chan, M. J. Addis, M. J. Boniface, P. Grimwood, A. Stevenson, C. Lahanier, and J. Stevenson. An Integrated Content and Metadata Based Retrieval System for Art. *IEEE Transactions on Image Processing*, 13(3): 302–313, March 2004.

J. Li and J. Z. Wang. Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models. *IEEE Transactions on Image Processing.*, 13(3):340–353, March 2004.

J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

T. Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2):79–116, November 1998.

S. P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, March 1982.

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

C. Lu, L. J. Latecki, N. Adluru, X. Yang, and H. Ling. Shape Guided Contour Grouping with Particle Filters. In *IEEE International Conference on Computer Vision*, October 2009.

M. Macri and M. Looper. *The New Catalog of Maya Hieroglyphs*, volume 1 The Classic Period Inscriptions. University of Oklahoma Press : Norman, 2003.

J. Mairal, G. Sapiro, and M. Elad. Learning Multiscale Sparse Representations for Image and Video Restoration. *Multiscale Modeling Simulation*, 7(1):214–241, April 2008.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11:19–60, March 2010.

S. Mallat and Z. Zhang. Matching Pursuits with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

F. Mendels, P. Vandergheynst, and J.-Ph. Thiran. Matching Pursuit-Based Shape Representation and Recognition Using Scale-Space. *International Journal of Imaging Systems and Technology*, 6(15):162–180, March 2006.

K. Mikolajczyk and C. Schmid. An Affine Invariant Interest Point Detector. In *Proceedings of the 7th European Conference on Computer Vision*, May 2002.

K. Mikolajczyk and C. Schmid. Scale and Affine Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

G. Mori, S. Belongie, and J. Malik. Efficient Shape Matching Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, November 2005.

R. Nock and F. Nielsen. On Weighting Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1223–1235, 2006.

B. A. Olshausen and D. J. Field. Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381:607–609, June 1996.

B. A. Olshausen and D. J. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1. *Vision Research*, 37:3311–3325, ??? 1997.

A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *Proceedings of the 9th European Conference on Computer Vision*, May 2006.

N. Payet and S. Todorovic. From Contours to 3D Object Detection and Pose Estimation. In *IEEE International Conference on Computer Vision*, November 2011.

M. Pitts and L. Matson. Writing in Maya Glyphs, 2008. URL http://www.famsi.org.

P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling Scenes with Local Descriptors and Latent Aspects. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2005.

P. Quelhas, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A Thousand Words in a Scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, September 2007.

M. Ranzato and Y. LeCun. A Sparse and Locally Shift Invariant Feature Extractor Applied to Document Images. In *Proceedings of the International Conference on Document Analysis and Recognition*, September 2007.

R. P. N. Rao. Probabilistic Analysis of an Ancient Undeciphered Script. *IEEE Computer*, 43(4): 76–80, April 2010.

R. Rigamonti, M. A. Brown, and V. Lepetit. Are Sparse Representations Really Relevant for Image Classification? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2011.

E. Roman-Rangel, C. Pallan, J.-M. Odobez, and D. Gatica-Perez. Analyzing Maya Hieroglyphs with Shape Context. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2009.

E. Roman-Rangel, C. Pallan, J.-M. Odobez, and D. Gatica-Perez. Searching the Past: An Improved Shape Descriptor to Retrieve Maya Hieroglyphs. In *Proceedings of the ACM International Conference in Multimedia*, November 2011a.

E. Roman-Rangel, C. Pallan, J.-M. Odobez, and D. Gatica-Perez. 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, August 2011b.

E. Roman-Rangel, J.-M. Odobez, and D. Gatica-Perez. Assessing Sparse Coding Methods for Contextual Shape Indexing of Maya Hieroglyphs. *Journal of Multimedia*, 7(2):179–192, April 2012.

B. C. Russell, J. Sivic, J. Ponce, and H. Dessales. Automatic Alignment of Paintings and Photographs Depicting a 3D Scene. In *Proceedings of the IEEE International Conference on Computer Vision, Workshop on 3D Representation for Recognition*, October 2011.

R. Sharer. *Daily Life in Maya Civilization.* Westport, CT: Greenwood, 1996.

J. Shotton, A. Blake, and R. Cipolla. Contour-Based Learning for Object Detection. In *IEEE International Conference on Computer Vision*, October 2005.

J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2003.

A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

P. Srinivasan, Q. Zhu, and J. Shi. Many-to-one contour matching for describing and discriminating object shape. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

F. Stanco, S. Battiato, and G. Gall. *Digital Imaging for Cultural Heritage Preservation: Analysis, Restoration, and Reconstruction of Ancient Artworks.* CRC Press, 2011.

D. S. Stuart, B. MacLeod, S. Martin, and Y. Polyukhovich. Glyphs on Pots: Decoding Classic Maya Ceramics. In *Sourcebook for the 29th Maya Hieroglyphic Forum*, 2005.

H. Sundar, D. Silver, N. Gagvani, and S. Dickinson. Skeleton Based Shape Matching and Retrieval. In *Proceedings of the Shape Modeling International*, May 2003.

R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer-Verlag New York, Inc., 1st edition, 2010.

C. H. Teh and R. T. Chin. On Image Analysis by the Method of Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1988.

J. E. S. Thompson. *A Catalog of Maya Hieroglyphs.* Norman : University of Oklahoma Press, 1962.

J. A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis. A Complete Optical Character Recognition Methodology for Historical Documents. In *Proceedings of the Eighth IAPR International Workshop on Document Analysis Systems*, 2008.

P. A. Viola and M. J. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2001.

C. Wang, J. Zhang, B. Yang, and L. Zhang. Sketch2Cartoon: Composing Cartoon Images by Sketching. In *Proceedings of the ACM International Conference in Multimedia*, November 2011a.

K. Wang, H.G. Zhang, L.S. Chai, Haiying, and Z.L. Ping. A Comparative Study of Moment-Based Shape Descriptors for Product Image Retrieval. In *International Conference on Image Analysis and Signal Processing*, October 2011b.

O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform Deblurring for Shaken Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and Shuicheng Yan. Sparse Representation for Computer Vision and Pattern Recognition. In *Proceedings of the IEEE*, March 2010.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2009a.

M. Yang, K. Kpalma, and J. Ronsin. A Survey of Shape Feature Extraction Techniques. *Pattern Recognition Techniques, Technology and Applications*, pages 43–90, 2008.

X. Yang, S. Koknar-Tezel, and L.J. Latecki. Locally Constrained Diffusion Process on Locally Densified Distance Spaces with Applications to Shape Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2009b.

C.T. Zahn and R.Z. Roskies. Fourier Descriptors for Plane Close Curves. *IEEE Transactions on Computers*, 21(3):269–281, 1972.

S. Zambanini and M. Kampel. Automatic Coin Classification by Image Matching. In *International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, 2011.

D. Zhang and G. Lu. Review of Shape Representation and Description Techniques. *Pattern Recognition*, 37(1):1–19, 2004.

Q. Zhu, L. Wang, Y. Wu, and J. Shi. Contour Context Selection for Object Detection: A Set-to-Set Contour Matching Approach. In *European Conference on Computer Vision*, 2008.

Y. Zhuang, Y. Zhuang, Q. Li, and L. Chen. Interactive High-dimensional Index for Large Chinese Calligraphic Character Databases. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(2), September 2007.

# Curriculum Vitae – Edgar F. Roman-Rangel

## Personal

Edgar F. ROMAN-RANGEL.
Idiap Research Institute
Centre du Parc. Rue Marconi 19.
PO Box 592. 1920 Martigny, Switzerland.
tel: +41 27 721 7775.
edgar.roman@idiap.ch
www.idiap.ch/∼eroman

## Education

**4th year PhD student** (Aug. 2008 - Present).
Thesis: *Statistical Shape Descriptors for Ancient Maya Hieroglyphs Analysis.*
Thesis director: Daniel Gatica-Perez. Co-director: Jean-Marc Odobez.
École Polytechnique Fédérale de Lausanne (EPFL, Swiss Federal Institute of Technology of Lausanne).
Submitted, and to be presented on 28th November, 2012.

**M.S. Computer Science** (Aug. 2004 - Oct. 2006).
Thesis (Combinatorial Optimization): *Analytical Tuning for the Cooling Scheme of the Simulated Annealing Algorithm Applied to the Protein Folding Problem.*
Thesis director: Juan Frausto Solis.
Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico (ITESM, Monterrey Institute of Technology and Higher Education).

**B.S. Computer Engineering** (Aug. 1999 - Jul. 2002). Universidad Morelos de Cuernavaca, Mexico (UMC, University Morelos of Cuernavaca).

## Work experience

**2008 August → Present** *Research assistant* as part of the Social Computing Group at Idiap Research Institute, Switzerland (www.idiap.ch, Funded by the Swiss National Science Foundation, SNSF).
Working in the project CODICES, developing techniques for semi-automatic processing of *Cultural and Historical Image Collections*, using Computer Vision, Image Retrieval, and Machine Learning techniques. More precisely, I have designed and developed a new shape descriptor for hieroglyphs of the Maya culture from ancient Mesoamerica.
This project maintains collaboration with the National Institute of Anthropology and History of Mexico (INAH, www.inah.gob.mx).
*Experience on:* shape-based image similarity for retrieval, machine learning techniques, topic models, and sparse coding methods.

**2012 June → 2012 September** *R+D Internship* at Microsoft Research Asia, Beijing. Working with the *Web Search and Mining Group* in the assessment of the retrieval performance of state-of-the-art shape descriptors for Oracle Bones Inscriptions (ancient Chinese characters). Research on the *characteristic scale* for Shape Context-like descriptors. Development of a visual retrieval system.
*Experience on:* Interest points, characteristic scales. C# and XAML.

**2007 November → 2008 April** *Project manager* at the National Bank of Mexico (Banamex, www.banamex.com). Leading the documentation of projects, and ensuring on-time delivery of several developed modules.

**2005 June → 2007 October** *Decision-making consultant* at Decisions and Logistics. Consulting for HSBC-Mexico in the CRM and the Logistics departments, performing mathematical analysis in the behavior of customers; implementing regression models to detect irregular transactions; designing mathematical models to minimize the costs of carrying and storing values and to maximize the level of service in ATM's.
*Experience on:* prediction with regression models.

**2004 April → 2004 October** *R+D Internship* at the Electric Research Institute, Mexico (IIE, www.iie.org.mx). Development of an information management system for 3-D models of oil platforms of the Gulf of Mexico.
*Experience on:* Programmable Macro Language from AVEVA Group plc.

**2002 April → 2002 October** *R+D Internship* at the IIE. Development of a graphic interface for the simulator of a nuclear power plant (Laguna Verde).
*Experience on:* Graphic programming with LabView.

### Awards and scholarships

- *Best Poster Award* at ENS/INRIA Visual Recognition and Machine Learning Summer School. Paris, 2011.
- *Idiap PhD Student Paper Award 2010.* This award is granted once a year to students being first authors in outstanding publications.
- ITESM Research Scholarship holder. During M.S. studies.
- Telmex foundation Scholarship holder. During B.S. and M.S. studies.
- University Morelos of Cuernavaca Scholarship holder. During B.S. studies.

### Programming skills

*Proficient with:* Matlab, C/C++, OpenCV, C#.
*Previous experience:* Java, Shell script, HTML, OpenGL, LabView, Python, XAML.

### Relevant coursework

*Scholar:*

- EPFL – Machine Learning.
- EPFL – Cognitive Vision for Cognitive Systems.
- ITESM – Search, Optimization, and Learning.
- ITESM – Artificial Intelligence.

*Others:*

- 2011 Visual Recognition and Machine Learning Summer School, ENS/INRIA. Paris, France.
- 2011 Hands on: Programming Python.
- 2007 Diploma: Programming Java (J2SE & J2EE) with Oracle 10g.
- 2006 XI Workshop on Mathematical Modeling. Department of Mathematics at the Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav). Mexico City, Mexico.

- 2005 2nd Leadership workshop at La Salle University. Cuernavaca, Mexico.

### Press

My research has been featured in scientific and mainstream media, including:

- Déchiffrer le passé avec du silicium. In *Horizons*, the Swiss magazine of scientific research, No 84, Mars 2010.
- Le logiciel qui a du caractère. In *Television Suisse Romande (TSR)*, June 4th, 2010.
- Maya script. In *Euronews*, June 22nd, 2010.

### Publications

**Journals**

Edgar Roman-Rangel, Jean-Marc Odobez, and Daniel Gatica-Perez. "Assessing Sparse Coding Methods for Contextual Shape Indexing of Maya Hieroglyphs". *Special Issue in Recent Achievements in Multimedia for Cultural Heritage. Journal of Multimedia.* 7(2):179–192. 2012.

Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. "Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptors". *International Journal of Computer Vision (IJCV), Special Issue in Cultural Heritage and Art Preservation.* 94(1):101–117. 2011.

**Conferences and workshops:**

Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. "Searching the Past: An Improved Shape Descriptor to Retrieve Maya Hieroglyphs". (full paper), *In Proc. ACM International Conference in Multimedia (MM).* Scottsdale, USA. November - December, 2011.

Daniel Gatica-Perez, Edgar Roman-Rangel, Jean-Marc Odobez, and Carlos Pallan. "New world, New Worlds: Visual Analysis of Pre-Columbian Pictorial Collections". *In Proc. International Workshop on Multimedia for Cultural Heritage (MM4CH).* Modena, Italy. April, 2011.

Edgar Roman-Rangel, Carlos Pallan, Jean-Marc Odobez, and Daniel Gatica-Perez. "Retrieving Ancient Maya Glyphs with Shape Context". *In Proc. IEEE International Conference on Computer Vision (ICCV), Workshop on eHeritage and Digital Art Preservation.* Kyoto, Japan. September - October, 2009.

Juan Frausto-Solis, E. F. Roman, David Romero, Xavier Soberon, Ernesto Liñán-García. "Analytically Tuned Simulated Annealing Applied to the Protein Folding Problem". *In Proc. International Conference on Computational Science.* Beijing, China. 2007.

### Activities

I have been invited to serve as part of Program and Technical Committees for scientific conferences and journals:

- 5th Mexican Conference on Pattern Recognition 2013
- ACM Multimedia 2012
- IEEE Transactions on Multimedia 2012

- Journal of Multimedia 2012
- IEEE ICCV 2011
- ACM ICMR 2011
- ACM Multimedia 2011
- ACM Multimedia 2010

References Available under request.

Personal interests

1. History, mythology, and legends.
2. Photography and cinema.
3. Biking, swimming, and traveling.