# Literature Synthesis: Bleek & Lloyd Dictionaries Project

Kyle Williams
Supervisor: H. Suleman

06 May 2009

## Abstract

Cultural heritage preservation is undertaken by museums and libraries worldwide with the goal of preserving historical artifacts and making them accessible. In many cases the historical artifacts are documents and therefore are best suited for preservation in a digital collection. Stories from the South Africa Bushman people exist in a digital collection called the Bleek & Lloyd collection. There are newly scanned Bushman dictionaries which can assist in providing translations and key understandings into the meanings of the words in these stories and the stories as a whole. However, because the stories and dictionaries are in digital collections there needs to be some sort of interaction between the two in order to facilitate this translation and understanding. The stories and dictionaries are stored in collections of scanned images and therefore image matching is well suited to providing interaction between the two. This paper explores previous work done in preserving cultural heritage, image matching and word spotting and shows that word spotting is well suited for matching images of hand written text in historical documents.

## 1 Introduction

The Bleek and Lloyd dictionaries are a collection of historical documents, which contains translations between English words and a variety of South African Bushman languages. The English-—Xam Bleek and Lloyd dictionary contains over 30 000 scanned images of translations between English words and words in the —Xam language. Storing and arranging of this collection of scanned images would make it useful and usable. Doing this involves archiving the material, making it searchable, presenting it in meaningful ways and providing for interaction between existing Bleek and Lloyd material [11] and the newly scanned dictionary.

A scanned collection of Bushman stories currently exists in a usable form [17]. The interaction mentioned above involves matching individual words in these stories to their corresponding words in the dictionaries, using content based image retrieval (CBIR) [2], and a subset of CBIR called word spotting [13].

This paper considers previous work done in CBIR and word spotting in order to assess ways in these methods can be used in preserving cultural heritage.

## 2 Bleek & Lloyd Collection

Suleman [17] created a usable digital archive of 14128 images, from a total of 157 notebooks from the Bleek & Lloyd collection. An XML-centric solution was chosen for this archive because it:

1. required no installation from the end user
2. was platform independent
3. allowed for easier processing
4. had long term preservation benefits

Suleman was able to show that the XML-centric approach did have many advantages over traditional database-based archives. However, scalability issues were identified as a limiting factor [17].

The resulting product of the work done by Suleman is available online [20] and in book and DVD format [15].

The work to be done in this project involves extending the digital Bleek & Lloyd collection to include the newly scanned dictionaries and also providing for interaction between the two.

# 3 Preserving Cultural Heritage

Museums and libraries world-wide digitise their invaluable historical documents with the goals of long term preservation and ease of access. A key requirement for digital collections is that objects are annotated in order for them to be accessible and exploitable. Annotation needs to be done either manually or automatically, however, manual annotation suffers from a key problem in that it is extremely tedious and expensive. In response to this, automatic systems have been built for storing, accessing and annotating digital collections [3]. These systems for managing digital collections have been used in a wide variety of projects for preserving cultural heritage [1, 21, 10].

## 3.1 The MEMORIAL Project

The MEMORIAL Project is a project funded by the European Union and undertaken by a multinational consortium. The goal of the memorial project is to enable the virtual distribution of paper-based archives which are currently held at museums and libraries [1]. The official title of the project is: "A Digital Document Workbench for Preservation of Personal Records in Virtual Memorials," which Antonacopoulos et al [1] note suggests that the project's focus is on information about people.

Antonacopoulos et al [1] document the use of the MEMORIAL project to create a digital collection of World War II personal records which contain information about people at Nazi run concentration camps. Antonacopoulos et al discuss how the MEMORIAL project framework was used for document input, image analysis, optical character recognition (OCR) and the creation of a Web-based portal which would allow users to access and make use of the content. In their concluding remarks, Antonacopoulos et al note that the MEMORIAL project was still underway, however the project's website appears to have been taken down. Regardless of this fact, the MEMORIAL project is/was a good example of making use of digital collections to preserve cultural heritage and make it usable and accessible.

## 3.2 The MICHAEL Project

The Multilingual Inventory of Cultural Heritage in Europe (MICHAEL) Project is a project funded by the European Commission to establish a new service for European cultural heritage. The project's vision is to create a service which will allow people to find and explore digital European cultural heritage on the Internet [21]..

## 3.3 Greek Orthodox Archdiocese of America

The Greek Orthodox Archdiocese of America (GOA) undertook a project to digitise their large collection of religious and historical artifacts. As is the case with most digital collections of cultural heritage, the goal of the project was to preserve the artifacts as well as make them accessible by users [10]. The GOA archive however was significantly different from other digital collections of cultural heritage in that the GOA had the need to make use of digital rights management (DRM) in order to protect and control access to the archive [10].

# 4 Content Based Image Retrieval

Content based image retrieval (CBIR) came about as a result of two fundamental shortcomings of text-based image retrieval: the amount of effort required to annotate images in large databases, and the subjectivity of human perception to the meaning of images [14].

There are generally three types of CBIR: primitive queries (query by example), semantic retrieval and automatic retrieval [4]. Primitive queries, or queries by example, are the most common types of queries in CBIR and therefore they are the only ones which will be discussed in full detail here, however, for completeness, the others will be discussed briefly.

In semantic retrieval, images are analysed and a set of possible interpretations are derived, each having some probability of being the correct meaning. Further semantic retrieval involves user feedback, which allows the system to learn about primitive features based on semantic concepts [4]. There are

few real life examples of automatic retrieval, however, one of its applications involves a CBIR system analysing colour in an image, determining whether the colours are "hot" or "cold", and returning images which convey a similar "mood" or "feeling" [4].

In primitive CBIR, images are analysed based on a number of primitive features, most notably colour, texture, shape and colour layout. This analysis usually takes place on segmented parts of the image, as it has been shown that the shape and colour layout analyses depend on good segmentation[14].

Once these features have been extracted from an image, it is possible to construct a signature for the image which can then be compared to signatures of other images in order to find matches.

### 4.1 Segmentation

In CBIR, images are often divided into parts, and then the features of each part are analysed separately. The goal of segmentation is to have more selective features in pixels, rather than more information about the image as a whole[16]. Smeulders et al [16] differentiate between four main types of segmentation:

- strong segmentation, in which a segment contains the pixels of an object in the real world and nothing else
- weak segmentation, in which segments contain data which is homogeneous according to some criterion
- signing, in which an object has a nearly fixed shape and a semantic meaning
- partitioning, in which a partition is simply a division of the data array

As noted by Smeulders et al, the different types of segmentation allow for different features to be extracted from an image[16].

### 4.2 Colour

Colour is one of the most widely used features in CBIR, as it is relatively robust to background complication and independent of image size and orientation [14]. In CBIR, the colour histogram is the most commonly used colour feature representation, as it shows the intensity of each of the three main colour channels [14]. These histograms allow for images to be compared, based on the similarity of their colour distributions. The colour histogram does however have one significant shortcoming in that it only shows the colour distribution of an image. Due to this, it could compute similar values for very different images if their colour distributions are similar [12]. Rasheed et al suggest an alternative in the form of a colour correlogram, which not only shows the intensity of the colour distributions, but also the spatial information of pixels in the image[12]. The approach of Rasheen et al. allows for significantly more meaning to be extracted from the colour in an image than the traditional colour histogram approach.

### 4.3 Texture

In CBIR, texture refers to the repetitive patterns which appear on the surfaces of images [14, 2]. Textures provide a further basis of comparison for two images, as the images can be compared based on the features of their textures. Textures are often domain specific, such as the textures of aerial imagery and medical imagery. It has been shown that texture features can be extracted by a variety of methods such as the transformation of the original pixels of an image, or wavelet transformations [2].

### 4.4 Shape

Shape is another basis upon which two images can be compared. A large variety of ways of detecting shapes for comparison have been suggested. Examples are: making use of salient edges [6], Fourier descriptors [23] and making use of image properties which are indirectly related to shapes, rather than the shapes themselves [5].

### 4.5 Signatures

The signature of an image is created based on its features. Datta et al [2] identify three main types of signatures:

1. a feature vector in which a single vector is used to describe the whole image (global)

2. a region-based signature in which each region is described by a separate vector (local)

3. a summary of local feature vectors

A local signature represents the specific details of an image and a global signature represents an image's "bigger picture" [2]. According to Datta et al, in recent years there has been a shift from global signatures towards local signatures. The reason for this shift is based on the growing understanding that local features often correspond with more meaningful aspects of an image [2], and these more meaningful aspects are more useful in CBIR.

## 4.6 Comparing Signatures

Signatures are compared based on some distance formula. The smaller the distance between images, the more similar they are. In this sense, images can be matched based on the similarity of their signatures. Datta et al [2] note that it becomes significantly more difficult to match signatures when region based (local) signatures have been used, due to the complexity of calculating the distances between the set of vectors.

## 4.7 CBIR Systems

### 4.7.1 retrievr & imgSeek

retrievr [22] and imgSeek [19] are two image-based search engines based on the fast multiresolution image querying algorithm developed by Jacobs et al [8]. The fast multiresolution image querying algorithm, and thus retrievr and imgSeek, uses a hand drawn sketch or low quality scan of the image to be retrieved. retrievr makes use of the hand drawn sketch or low quality scan to search Flickr [18] for similar photos, whereas imgSeek is a photo collection manager with built in CBIR. Jacobs et al. [8] tested their algorithm by using sketches and low quality scans of actual images to see if they could find the correct image in a database of sample images. They found that their algorithm was extremely fast and effective and able to pinpoint the correct image to within a 1% subset of the original sample - that is, if there were 100 images in the sample, they were able to find the correct image almost every time. However, for large sample sizes, they were able to find 1% of the sample size that were possible matches.

### 4.7.2 IBM's Query By Image Content System

IBM's Query By Image Content (QBIC) system is another CBIR system which performs searches based on the primitive features of an input image [7]. The system enhances the quality of its searches by allowing the user to provide key words in addition to the image used for searching. IBM has identified several real world uses of the QBIC system such as using it to search clothing catalogues to find certain styles of clothes [7].

# 5 Word Spotting

Word spotting is a technique for grouping occurrences of the same word, where it exists in multiple locations in a document. It is based on the user providing a key word, and then image matching is done to try and find other occurrences of that word [13]. Word spotting can be considered a subset of CBIR because it is based on comparing the distances between two images - the key image and other candidate occurrences of the key image. Distances are calculated based on the differences between the feature vectors or profiles of words [13, 9].

Word spotting is well suited to the case of handwritten historical documents, where optical character recognition (OCR) techniques do not work well [9]. Rath et al. show that word spotting can be successfully applied to historical documents, by demonstrating its use on the George Washington collection. It was shown that a word's profile can be created based on a few key features.

Figure 1 shows the original word Rath et al [13] were working with.

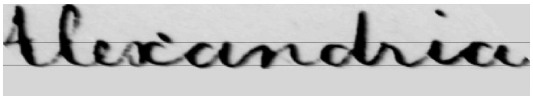

Figure 1: Original image

The first feature captured by Rath et al [13] was the distribution of ink along one of the two dimensions of an image, referred to as the projection profile (Figure 2).

The next features captured were the upper and lower profiles (shapes) of a word (Figure 3). These

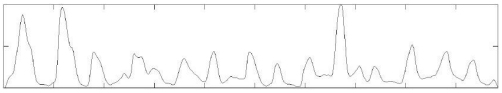

Figure 2: Projection profile

features were captured by applying background-foreground separation techniques.

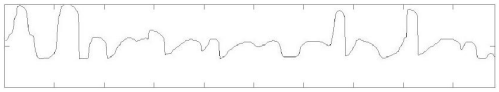

Figure 3: Upper profile

The last feature extracted from a word was information about its "inner" content (Figure 4). This information was derived based on the number of background to text transitions which took place in an image.

Figure 4: Background-foreground transitions

The information extracted by Rath et al. [13] gave them three key pieces of information about a word:

1. The amount of ink used in an word
2. The upper profile and lower profile of a word
3. An idea about the inner structure of a word

Armed with this information, they apply a dynamic time warping (DTW) algorithm for image matching. DTW is used because it allows for writing variation as it makes use of a common time axis. It is shown that DTW greatly outperforms other distance measuring techniques such as Euclidean Distance Matching [13].

Furthermore, Rath et al. clearly show that word spotting is a viable and practical method for matching images containing handwritten text.

Leydier et al. [9] show the application of word spotting to medieval manuscript images [9]. The key idea revealed by the work done by Leydier et al. is that of domain specific word spotting, that is, they base the feature vector for each word on specific attributes of written medieval words. Their proposed approach involves no layout segmentation, no binarisation of images, and tolerates low quality and distorted images. Instead of focusing on the greyscale transformation of an image, the focus is instead placed on informative parts of an image, which are the most discriminant parts of an image and referred to as zones of interest (ZOIs) [9]. They found that the orientation of the gradients of strokes at these ZOIs is the most efficient description of a word's shape. They furthermore showed that focusing on other features gave poor results [9].

# 6 Comparison of CBIR and Word Spotting

It has been shown that CBIR and word spotting can both be used successfully for image matching. The table below provides a brief comparison between the two techniques.

| | **CBIR** | **Word Spotting** |
|---|---|---|
| **Focus** | Pictures | Handwritten Words |
| **Primitive features** | Colour, texture, shape | Projection Profile, word profile, background-foreground transitions |
| **Applications** | Image searching | Word matching |

Table 1: Comparison of CBIR and word spotting

As is evident from the table, word spotting is a lot better suited to the task of providing interaction between words in the current Bleek & Lloyd collection and the newly scanned dictionary than traditional CBIR. This is because word spotting is a subset of CBIR which specifically focuses on matching images of written words, and therefore pays special attention to the features of those types of images. In their studies, Rath et al. [13] and Leydier et al. [9] clearly show how word spotting techniques can be applied to image matching for written words in a simple and straight-forward manner.

# 7 Conclusion

It has been shown that CBIR is largely based on extracting key features from images, deriving a signature, and then comparing that signature to the signature of other images, based on some measurement of similarity. Furthermore, CBIR has been successfully deployed in theoretical and commercial applications. It was shown however, that a subset of CBIR called word spotting is a lot better suited to the task of matching images of hand written words. This is due to the fact that word spotting pays special attention to the features of images of hand written words, whereas traditional CBIR is more focused on pictures. A viable and practical method for word spotting has been explored [13], as well as an application of word spotting to a specific domain [9]. This shows that word spotting is a suitable and viable technique to be used for the required interaction between the Bleek & Lloyd stories and dictionaries.

# References


[1] Apostolos Antonacopoulos and Dimosthenis Karatzas. Document image analysis for world war ii personal records. In *International Workshop on Document Image Analysis for Libraries (DIAL2004)*, pages 336–341. IEEE-CS Press, 2004.

[2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[3] Reim Doumat, Elöd Egyed-Zsigmond, Jean-Marie Pinon, and Emese Csiszar. Online ancient documents: Armarius. In *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, pages 127–130, New York, NY, USA, 2008. ACM.

[4] J. Eakins and M. Graham. Content-based image retrieval. Technical report, JTAP, 1999. JISC Technology Applications Programme Report 39.

[5] George Gagaudakis and Paul L. Rosin. Incorporating shape into histograms for cbir. *Pattern Recognition*, 35(1):81–91, 2002.

[6] Jun Wei Han and Lei Guo. A shape-based image retrieval method using salient edges. *Signal Processing: Image Communication*, 18(2):141–156, 2003.

[7] IBM. Query by image content (qbic). *http://domino.research.ibm.com/comm/pr.nsf/pages/rsc.qbic.h* 2009. Accessed: 06 May 2009.

[8] Charles E. Jacobs, Adam Finkelstein, and David H. Salesin. Fast multiresolution image querying. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 277–286, New York, NY, USA, 1995. ACM.

[9] Yann Leydier, Frank Lebourgeois, and Hubert Emptoz. Text search for medieval manuscript images. *Pattern Recognition*, 40(12):3552–3567, 2007.

[10] Theo Nicolakis, Carlos E. Pizano, Bianca Prumo, and Mitchell Webb. Protecting digital archives at the greek orthodox archdiocese of america. In *DRM '03: Proceedings of the 3rd ACM workshop on Digital rights management*, pages 13–26, New York, NY, USA, 2003. ACM.

[11] J. Parkington. Conference report - bleek and lloyd 1870-1991. *Social Dynamics-a Journal of the Centre for African Studies University of Cape Town*, 17(2):168–170, DEC 1991. PT: J; NR: 0; TC: 0; J9: SOC DYNAMICS; PG: 3; GA: HE868.

[12] Waqas Rasheed, Youngeun An, Sungbum Pan, Ilhoe Jeong, Jongan Park, and Jinsuk Kang. Image retrieval using maximum frequency of local histogram based color correlogram. In *AMS '08: Proceedings of the 2008 Second Asia International Conference on Modelling & Simulation (AMS)*, pages 322–326, Washington, DC, USA, 2008. IEEE Computer Society.

[13] Tony M. Rath and R. Manmatha. Word spotting for historical documents. *International Journal on Document Analysis and Recognition*, 9(2-4):139–152, APR 2007. PT: J; NR: 42; TC: 5; J9: INT J DOC ANAL RECOGNIT; PG: 14; GA: 185UE.

[14] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, 1999.

[15] P. Skotnes. *Claim to the Country: the archive of Wilhelm Bleek and Lucy Lloyd*. Jacana Media and Ohio University Press, 2007.

[16] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[17] H. Suleman. Digital libraries without databases: The bleek and lloyd collection. In *ECDL*, pages 392–403, 2007. crossref: DBLP:conf/ercimdl/2007.

[18] Various. Flickr. *http://www.flickr.com/*, 2009. Accessed: 06 May 2009.

[19] Various. imgseek. *http://www.imgseek.net/*, 2009. Accessed: 06 May 2009.

[20] Various. Lloyd bleek project. *http://www.lloydbleekcollection.uct.ac.za/index.jsp*, 2009. Accessed: 02 May 2009.

[21] Various. The michael project. *http://www.michael-culture.org/en/about/project*, 2009. Accessed: 06 May 2009.

[22] Various. retrievr. *http://labs.systemone.at/retrievr*, 2009. Accessed: 06 May 2009.

[23] Dengsheng Zhang and Guojun Lu. Shape-based image retrieval using generic fourier descriptor. *Signal Processing: Image Communication*, 17(10):825–848, 2002.