Machine Vision and Applications
DOI 10.1007/s00138-016-0819-x

ORIGINAL PAPER

Image-based search and retrieval for biface artefacts using features capturing archaeologically significant characteristics

Mark Eramian[1] · **Ekta Walia**[1] · **Christopher Power**[2] · **Paul Cairns**[2] · **Andrew Lewis**[2]

Received: 13 April 2016 / Revised: 6 November 2016 / Accepted: 15 November 2016

Abstract Archaeologists are currently producing huge numbers of digitized photographs to record and preserve artefact finds. These images are used to identify and categorize artefacts and reason about connections between artefacts and perform outreach to the public. However, finding specific types of images within collections remains a major challenge. Often, the metadata associated with images is sparse or is inconsistent. This makes keyword-based exploratory search difficult, leaving researchers to rely on serendipity and slowing down the research process. We present an image-based retrieval system that addresses this problem for biface artefacts. In order to identify artefact characteristics that need to be captured by image features, we conducted a contextual inquiry study with experts in bifaces. We then devised several descriptors for matching images of bifaces with similar artefacts. We evaluated the performance of these descriptors using measures that specifically look at the differences between the sets of images returned by the search system using different descriptors. Through this nuanced approach, we have provided a comprehensive analysis of the strengths and weaknesses of the different descriptors and identified implications for design in the search systems for archaeology.

Keywords Image retrieval · Image-based search · Biface · Archaeology · Artifacts · Flint

✉ Mark Eramian
mge314@mail.usask.ca

Ekta Walia
ewb178@mail.usask.ca

Christopher Power
christopher.power@york.ac.uk

Paul Cairns
paul.cairns@york.ac.uk

Andrew Lewis
andrew.lewis@york.ac.uk

[1] Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[2] Department of Computer Science, University of York, York, UK

1 Introduction

It has long been recognized that the process of archaeology is by necessity a destructive one. In 1908, Sayce [31] echoed the sentiment of Flinders Petrie when he commented that "Scientific excavation means, before all things else, careful observation and record of every piece of pottery, however apparently worthless, which the excavator disinters".

The very process which enables an archaeologist to understand each layer of an excavation requires the previous layer to be removed. The process of excavation is an unrepeatable operation [6]. For this reason, modern archaeology relies heavily on the use of recording technology to retain as much information as possible. One of the most common recording methods for in situ and excavated artefacts is that of digital photography.

Modern archaeological studies generate thousands of digital photographs of archaeological artefacts which are often assembled into large public archives such as those provided by the British Archaeology Data Service [1]. Archaeologists utilize such databases in identification and classification of newly discovered artefacts, training new archaeologists, and in public engagement with archaeological history. Currently, these archives are limited in their ability to allow archaeologists to search for relevant images related to a particular task. Search is often limited to keyword searches, or, at best, faceted browsing, which is problematic when image

Published online: 10 December 2016

metadata is incomplete or in situations where the user is unfamiliar with the metadata schemes that are used in a particular archive.

The *Digging into Archaeological Data: Image Search and Markup (DADAISM)* project [12] is aimed at developing advanced computing technology, including image processing, to provide more useful and usable interactive systems for archaeologists to work with image archives. This paper reports on a major component of this project: a content-based image retrieval system for images of biface artefacts. This system uses image features to match a query image to images in the database with similar properties. The image features were developed with input from research archaeologists so that the image features capture properties of bifaces that would be used during classification tasks. This paper reports on the collection of data from archaeologists, the development of the image-based search algorithm, and an evaluation study to validate the results against existing classifications already present in the image metadata.

1.1 Related work

The image analysis and pattern recognition literature contain several works related to feature extraction, similarity grouping, and classification of archaeological artefacts.

As part of Graphically Oriented Archaeological Database project, the authors in [21] presented an automatic method for shape extraction from raster images of line drawings to facilitate matching and retrieval. Smith et al. [32] developed a scheme to classify thin-shell ceramics based on colour and texture descriptors in order to aid in vessel reconstructions, using a new feature based on total variation geometry along with SIFT (scale-invariant feature transform). Abadi et al. [2] present a system for automatic texture characterization and classification of ceramic pastes, fabrics, and surfaces. They use Gabor filter along with linear discriminant analysis and k-nearest neighbour techniques in order to achieve the desired objectives. In [5], the authors analyse surface properties in pottery and lithic artefacts. They define texture with attributes such as coarseness, contrast, directionality, line-likeness, regularity, and roughness. Durham et al. [13] use generalized Hough transform as a practicable and robust tool for matching whole and partial artefact shapes.

Few examples of image-based identification systems for archaeological artefacts also exist in the literature. The work in [34] illustrates the use of computer vision techniques for the development of content-based image retrieval system for historic glass and an automatic system for mediaeval coin classification. In another work by [36] a prototype is presented which allows the end-user to search for similar digital library objects based on the image content. An integrated content and metadata-based retrieval system for art images in the domain of museum and gallery image collections is presented in [22]. Image retrieval methods are proposed to perform query on the basis of subimages. It also presents methods of querying by very low-quality images. IBISA [26] is a software tool that allows the user to perform image-based searches on the database of digital images of archaeological objects such as ancient coins. This system performs image segmentation with a method based on active contours and image registration with Fourier–Mellin Transform and computes similarity with classic intercorrelation factor. An extension of this system which is robust to lighting conditions [25] has also been proposed recently. CLAROS [10] allows image-based searches for sculpture and pottery images based on scale-, illumination-, and viewpoint-invariant image patches encoded using a bag-of-words scheme.

While the use of images, either hand drawings or photographs, remains the most common visual record of found lithics artefacts, there are opportunities to use 3D scanning technology to improve various aspects of archaeological work. For example, Grosman et al. [15] proposed the use of 3D scanning for purposes of documenting the essential characteristics of lithics in a standardized way that is less open to interpretation and error. Lin et al. [23] proposed the use of 3D scanning for analysis of lithic artefacts by using precise measurements of the cortex recorded in scans to make approximations of the size of artefacts. Finally, there are examples of people using 3D scanning for more specific types of analysis, such as Evans and Donahue [14] who looked at microwear on lithic artefacts.

Even with these advances in 3D scanning, there are currently still many thousands more photographs taken of artefacts than there are of 3D scans. Financial and time resource costs are one reason for this, as is the need to capture artefacts at different stages, including on site where scanners are unlikely to be available. In other cases, it is that the 3D scans provide some information, while images can provide different perspectives on the objects, such as the original patination on the surfaces of flint bifaces, and thus both are likely to be in ongoing use [8]. Combining this with the thousands of images that are already in archives worldwide, it is the case that archaeologists are going to need to find and work with images for the foreseeable future. The purposes for working with these images are varied; however, often it is to compare and identify objects to understand their function in society. These, and other purposes for using images in archaeology, are discussed in the contextual inquiry study described in Sect. 3.

With our work, we have expanded the corpus of image-based search of archaeological images to biface artefacts using image features based on established archaeological methodology for identifying similarities between the artefacts. Expert-verified metadata is used to validate the retrieval performance.

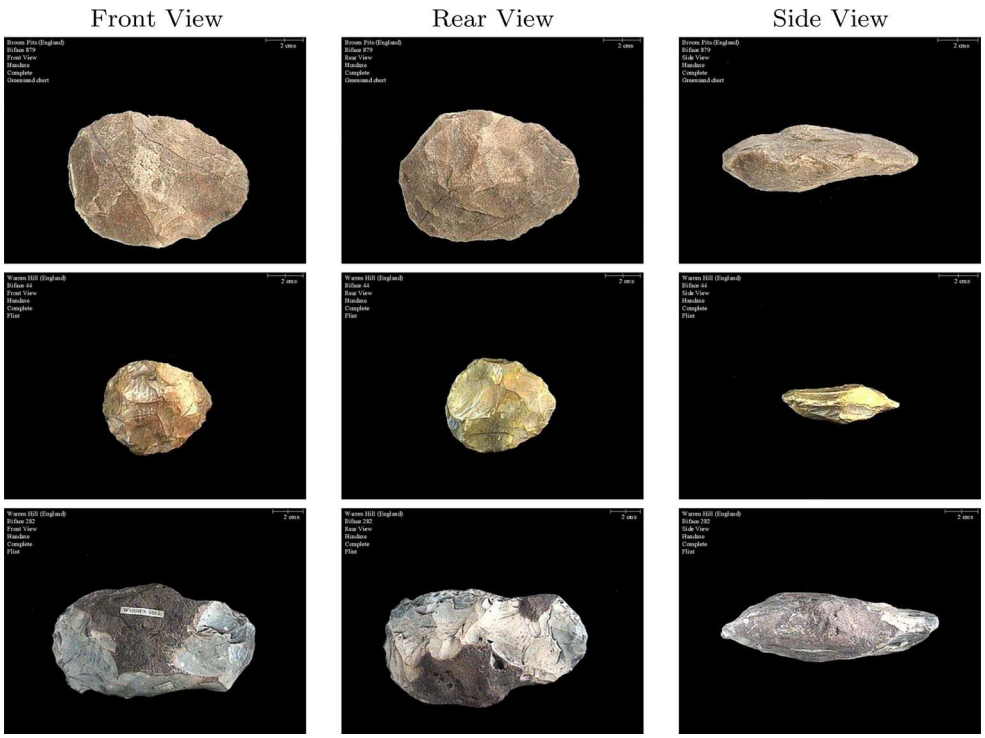


Fig. 1 Example front, rear, and side views of four bifaces. Each *row* contains different views of the same artefact

2 Materials and methods

2.1 Data set

Our data set is a public database from the British Archaeology Data Service consisting of 3501 images, each 500 × 375 pixels in size, of 1167 biface artefacts in three views: front, rear, and side [27]. Figure 1 contains several examples of bifaces of varying size, shape, and texture, with different views of the same artefact shown in the same row. The images were captured at varying resolutions indicated by the scale bar on the top-right corner of each image.

The data set also consists of 52 metadata fields for each artefact including the site and country from where the biface was excavated, the artefact's raw material, type, and its physical dimensions. The metadata was verified by biface experts and is of high quality, with little to no missing information. We utilized this information to define and measure relevance of results retrieved by our proposed system.

2.2 Image features for biface search

Image-based features suitable for representing the characteristics of bifaces that are important for identification and classification were selected after conducting a contextual inquiry (CI) study in which research archaeologists with relevant experience were interviewed. The CI study and its outcomes are described in Sect. 3. The specific features that were selected based on the result of CI study are described in Sects. 4.2 and 4.3.

2.3 Image-based search algorithm

An algorithm for identifying biface images similar to that of a query image based on the selected image features was developed. The details of this algorithm are presented in Sect. 5.

2.4 Evaluation of image-based search algorithm

The biface search algorithm was validated in several ways by comparing the metadata of the images returned by the search algorithm to the metadata of the query image and by looking at patterns of disagreements in the retrieval results for different image descriptors. The evaluation methodology and results of the evaluation are described in Sect. 6.

3 Biface classification contextual inquiry study

We conducted a contextual inquiry (CI) study to gain insight into how archaeologists judge similarity of bifaces. CI is a qualitative method developed by Beyer and Holtzblatt [7] to help design interactive systems that support users in their tasks. A CI study traditionally will have participants and researchers working in a collaborative way to explore different tasks *in situ*. In this case, we examined tasks where

researchers were using image archives for research purposes.

3.1 Participants

Four participants were recruited through mailing lists and personal contacts available to the Archaeological Data Service (ADS) at the University of York. Each participant had over 5 years of experience as a research archaeologist, made extensive use of picture archives in their work, and previously worked on prehistory sites with artefacts that included biface artefacts.

3.2 Materials

Interviewees were provided with an information sheet that discussed how the interview would proceed and were provided with an informed consent form. Interviews were recorded on a Panasonic HD video camera, and the interviewer took notes to support analysis of the recording. Of particular interest to the interviewer was the documentation of explicit task flows and tacit knowledge the participant was using to make decisions. After the interview was completed, participants were provided with a demographic questionnaire through the online questionnaire tool Qualtrics.

During the interview, participants were shown some preliminary results of the image processing algorithms to give them some insight into how their data would be used. These images were printed in full colour at 150 dots per inch (dpi) to ensure clarity in the presentation.

3.3 Procedure

Participants were met in a comfortable, quiet setting with a computer connected to the Internet. Participants were asked to bring with them any physical or digital materials that they use in typical, day-to-day work when trying to identify or classify artefacts.

Participants were asked about their own background in archaeology, their training, and their area of expertise. They were also asked what they considered to be the most important tasks that they undertook with online archaeological archives.

The CI study consisted of an initial semi-structured interview regarding how and when participants interacted with image archives and what they felt the most important aspects of the archives were. Further, the interview explored issues around favourite and least favourite aspects of working with image archives. This initial interview provided some initial context regarding what users do with their data.

Participants were asked about a recent time that they had used image archives to solve a research question and were asked to demonstrate how they used the archive to solve that problem. The participants demonstrated the tasks, describing how and why they used the archive. The interviewer would periodically stop the participant to discuss particular interesting aspects of the interaction, to elaborate on particular challenges or problems highlighted by the participant, or to clarify the purpose of particular actions. In this way, the interviewer was able to gain a better understanding of tacit aspects of the interaction that would not necessarily be raised during a traditional interview.

Finally, users were shown a set of images of bifaces and asked about different features that were used for classification and identification. This part of the interview led to a list of common features, as well as a set of research publications and secondary resources that would elaborate these features.

3.4 Results

All participants were trying to find information about artefacts that related to the ones they were already working with. However, the final uses of the images were varied. A few researchers were working on presentations and displays for public engagement, while others worked to find images for inclusion in publications or lectures. However, the majority of the participants were trying to collect together sets of similar artefacts in order to identify an artefact they had in hand or, in some cases, to pass off that identification task to another specialist team member.

In these cases, there were often several types of analysis that would be conducted by other team members in parallel, with individuals contributing analyses from a number of perspectives. For example, if the artefact is longer than they are wide, it indicates that the artefact was, perhaps, the end product, whereas other lithics in an assemblage were potentially flakes from the manufacturing process. If the blade is heavily worked, meaning much of the stone has had the cortex removed, it gives indication of the advancement of the civilization that created the artefact. Finally, morphology will often indicate the function of the tool, with the visual characteristics providing a means of classifying the tools by its function in society [3,28].

The recordings of the contextual inquiries were reviewed and partially transcribed to identify the tasks undertaken by participants in image archives, with a particular focus on the types of information they used to formulate and refine their queries. For purposes of this paper, we will focus on the characteristics of bifaces that emerged from this analysis. Experts reported that the following characteristics of biface tools are used to classify artefacts, judge the similarity of

artefacts, and to formulate new queries for image search systems:

- *Blade versus flake* The artefact will be classified as a blade or a flake. Blades are twice as long as they are wide; flakes are less than twice as long as they are wide.
- *Worked versus unworked* A distinction is made between the cortex and core of the object. The cortex is the surface of the stone which has been worked by geological processes; the core is the interior part and is desired for tool working. As civilization is advanced, manufacture moved to using more of the core, while older artefacts exhibit more cortex. Artefacts may be distinguished by their type of *removal*: primary (removed from nearly unworked stone with substantial cortex remaining), secondary (a biface with only a thin seam of cortex created from stone closer to the core), or tertiary (flakes that have been created from the deep interior of the stone after primary and secondary strikes have removed the cortex and therefore exhibit a smooth, glassy quality).
- *Retouched versus non-retouched* Once a removal has been struck from the core, the piece will either be left as rubbish (because it is small or brittle), picked up and used as is, or retouched. Retouched pieces are worked into a specific shape for a specific task, e.g. knapping to give it a sharp edge. The manner of retouching was reported to be of significant interest to archaeological research.
- *Colour* The colour of stone bifaces from the same region may or may not be of a consistent colour and therefore may or not be diagnostic. Whether colour is of diagnostic use may depend on the data set and geographic regions under consideration.
- *Morphology* Morphology refers to the shape of the biface, or the number, shape, and length of individual tines on the artefact. It is the most common method of matching finds to a geographic region or culture. Complications arise when some morphological differences are derived from other morphologies; for example, a tine could have broken off and retouched into another type of biface through knapping.

In the next section we describe how different characteristics are mapped to image descriptors to match images automatically.

4 Biface image descriptors

The CI study indicated that size, shape (morphology), and texture are diagnostic for bifaces. Our biface descriptor is a concatenation of image features capturing these three aspects. One set of morphological features capturing size and shape, and thirteen competing texture descriptors are described in the following sections. We begin with a description of the image preprocessing applied prior to extracting features.

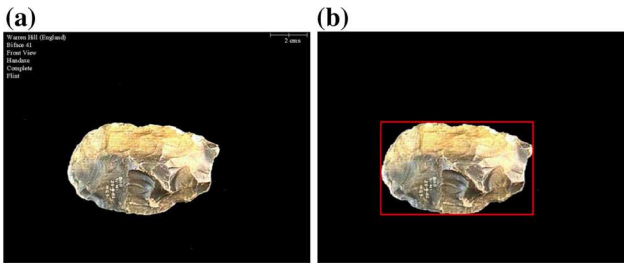


Fig. 2 Preprocessing of images. Annotations and scale are removed, and bounding box of artefact determined

4.1 Image preprocessing

Biface images were preprocessed to isolate the object of interest from the rest of the image. A sample unprocessed image is shown in Fig. 2. Images were smoothed by using a 5×5 Gaussian filter with $\sigma = 0.5$ pixels. Images were then converted to the HSV colour space, and the biface object is segmented by selecting all pixels that have neither minimum nor maximum saturation. This gives accurate segmentation results for almost all bifaces, except for a very few which have very dark or very bright areas on their surface. To address this problem, all holes in the detected foreground regions were filled. The bounding box of the segmented artefacts was determined from their segmentations as in Fig. 2. Finally the subimages defined by the bounding boxes were converted from RGB colour to greyscale (luminance as defined by BT.601 standard). All features comprising the biface descriptor are extracted from these greyscale images.

4.2 Morphological features

In this section we describe the size (geometric) and shape features that contribute to the biface descriptor.

4.2.1 Geometric features

Table 1 describes the geometric features extracted for a biface.

Scale normalization was achieved by determining the image's resolution in pixels per millimetre from the scale bar and converting distances and areas to units of mm and mm^2, respectively. We refer collectively to the features $\{g_1, g_2, \ldots, g_7\}$ as g.

Table 1 Geometric features computed for the biface descriptor

Feature	Description
g_1	Scale-normalized length of the artefact's bounding box
g_2	Scale-normalized width of the artefact's bounding box
g_3	Scale-normalized area (number of pixels in the artefact's segmentation)
g_4	The scale-normalized breadth of the artefact at 20% of the distance along the length of the bounding box
g_5	The scale-normalized breadth of the artefact at 80% of the distance along the length of the bounding box
g_6	The ratio g_2/g_1
g_7	The ratio g_4/g_5

4.2.2 Shape features

Biface shape is captured by a vector of Fourier descriptors. Following the centroid distance method of Zhang and Lu [38], the distance, $r(t)$, between the centroid of the artefact and the t-th perimeter boundary point was computed. These distances are then encoded using the discrete Fourier transform:

$$FD_l = \frac{1}{L}\sum_{t=0}^{L-1} r(t)\exp\left(\frac{-j2\pi lt}{L}\right), \quad l = 0, 1, \ldots, L-1. \tag{1}$$

where L is the number of boundary points. A Fourier descriptor FD_l of a given shape is translation invariant and is made rotation invariant by retaining only the magnitude and discarding the phase information. For scale invariance, the magnitude of FD_l is divided by the DC component. The DC-normalized magnitude of the first 39 Fourier coefficients was retained for the shape description. Thus, the vector of Fourier descriptor features encoding a biface's shape is:

$$f = \left\{\frac{|FD_1|}{|FD_0|}, \frac{|FD_2|}{|FD_0|}, \ldots, \frac{|FD_{39}|}{|FD_0|}.\right\} \tag{2}$$

4.3 Texture descriptors

The CI study indicated that texture of a biface's surface appearance was an important indicator of the biface's raw material type. In this section we describe 13 candidate texture descriptors. Some are existing features taken directly from existing literature, and others are new or are variations of existing features. The relative effectiveness of these 13 descriptors is evaluated in Sect. 6.

4.3.1 Uniform local binary patterns (ULBP)

Our ULBP descriptor consists of the 8-bit uniform local binary patterns as described in [33]. It is pertinent to mention here that we discarded the 59th bin containing non-uniform patterns as they have been observed to not be discriminating [33].

4.3.2 Orthogonal combination of linear binary patterns (OCLBP)

Our OCLBP descriptor consists of the features exactly as described by Zhu et al. [40].

4.3.3 Segmentation-based fractal texture analysis (SFTA)

Our SFTA descriptor uses segmentation-based fractal texture analysis [11] which decomposes the image into a set of binary images using a multithresholding scheme, and the fractal dimensions of the resulting regions describe the texture patterns. We used $n_t = 8$ thresholds in our implementation. We did not use size (pixel count) and mean greylevel as suggested in [11] for feature vector construction because they are not discriminative for our data set. We only used the fractal dimension computed over different binary images resulting from the binary decomposition algorithm. Thus, we generated $2n_t$ features using the box-counting algorithm for n_t thresholds, for a total of 16 features.

4.3.4 Global phase congruency histogram (GPCH)

Phase congruency (PC) is a measure of feature significance that is particularly robust to changes in illumination and contrast [19]. PC is based on the fact that the Fourier components are all in phase at the locations of significant features, such as step edges. It is a dimensionless quantity for measuring the consistency of local phase over different scales. It has been successfully used as feature in applications such as finger-knuckle-print recognition [39] and pose estimation for face recognition [30].

To compute a local PC map, a bank of quadrature-pair log-Gabor wavelets is applied to the image. Using the author's MATLAB implementation [20] of the method in [19], we used $n = 3$ scales of log-Gabor wavelet and 6 orientations $\theta = \{0, \frac{\pi}{6}, \frac{2\pi}{6}, \frac{3\pi}{6}, \frac{4\pi}{6}, \frac{5\pi}{6}\}$. We then computed a global phase congruency histogram (GPCH) by dividing the range of possible PC values, [0.0, 1.0], into 10 equal subintervals and counting the number of PC values in the PC map falling into each subinterval. The result is a vector of 10 texture features.

4.3.5 Angular radial phase congruency histogram (ARPCH)

We generated PC maps as in the previous section and subdivided the maps into radial sectors using the angular radial partitioning (ARP) described in [9]. The application of this partitioning scheme to PC maps is novel. ARP overlays on the PC maps P concentric circles which are divided into Q angular partitions. This gives $P \times Q$ sectors on the PC map.

A histogram of PC values is constructed for each sector of the PC map. We divided the range of possible PC values [0.0, 1.0] into B equal-sized subintervals and created a histogram H_s for each sector s of an image as follows:

$$H_s(b) = \frac{\text{count}_\text{b}(PC(s))}{\text{area}(s)}, \quad b = 1, 2, \ldots, B, \tag{3}$$

where B is the number of histogram bins, $PC(s)$ is the multiset of phase congruency values in the sector s, and $\text{count}_\text{b}(PC(s))$ is the number of elements in $PC(s)$ which fall into bin b.

We then construct vectors A_i by concatenating the values of the i-th bin of each sector histogram: $A_i = (H_1(i), H_2(i), \ldots, H_S(i))$. These are concatenated into a feature vector $A = (A_1, A_2, \ldots, A_B)$. We used $S = P \times Q = 3 \times 8 = 24$ sectors as shown in Fig. 3, and $B = 10$ histogram bins, resulting in 240 features in A.

To reduce size of this descriptor, A is subdivided into subvectors $\hat{A}_j$ of length $S/2$. We then perform singular value decomposition on each $\hat{A}_j$:

$$\hat{A}_{j(1\times S/2)} = U_{j(1\times k)} \Sigma_{j(k\times k)} V^T_{j(k\times S/2)}. \tag{4}$$

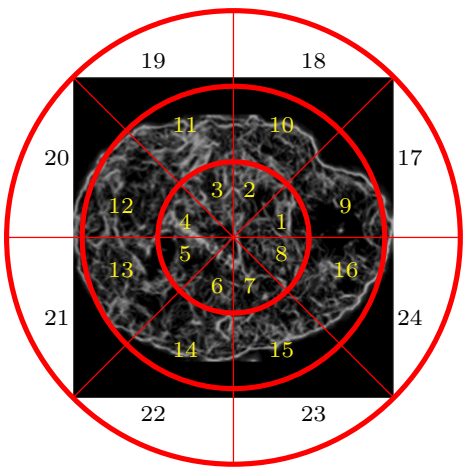


Fig. 3 The radial and angular divisions of the PC map for $P = 3$ and $Q = 8$. *Numbers* indicate the sector ordering used when concatenating sector histograms

Since $k = \min(1, S/2) = 1$, Σ_j is a 1×1 matrix, regardless of S, and we represent each subvector $\hat{A}_j$ by the singular value Σ_j. These are concatenated to form our 20-feature ARPCH descriptor of $(\Sigma_1, \Sigma_2, \ldots, \Sigma_{2B})$.

4.3.6 Orientation-based phase congruency histograms (OPCH)

In this novel variation of PC-histogram-based features, instead of computing one PC map, as with the GPCH and ARPCH features, we computed eight separate PC maps for different filter orientations using [20]. We computed 8 different PC maps for the set of orientations $\theta = \{0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}\}$ using $n = 4$ scales and took the summation over all scales n for each orientation in θ.

We then generated the histogram of PC values for each direction in θ:

$$H_\theta(b) = \frac{\text{count}_\text{b}(PC_\theta(I))}{\text{rows}(I) \cdot \text{cols}(I)}, \quad b = 1, 2, \ldots, B \tag{5}$$

where $\text{rows}(I)$ and $\text{cols}(I)$ are the dimensions of the bounding box found during image preprocessing. Again the range of PC values was divided into $B = 10$ equal-size subintervals. The histograms for each PC map were concatenated to form our orientation-based phase congruency histogram (OPCH) descriptor consisting of 80 features.

4.3.7 Gabor wavelet features (GWF)

We extracted features from the Gabor wavelet transform of a artefact's image as in [24]. A filter bank of Gabor wavelet functions with different orientations is convolved with the image to obtain a complex response $R_{s\alpha}(x, y)$ as follows

$$R_{s\alpha}(x, y) = \sum_m \sum_n I(x - m, y - n)\Psi^*_{s\alpha}(m, n) \tag{6}$$

where I is the image; $\Psi^*_{s\alpha}$ is the complex conjugate of $\Psi_{s\alpha}$, which is a self-similar function generated from the dilation and rotation of the mother wavelet Ψ. The definitions of Ψ, $\Psi_{s\alpha}$ and $\Psi^*_{s\alpha}$ can be found in Eqs. 1–3 of [24] where they are called g, G, and g_{mn}, respectively. Intuitively s determines the scale of the filter, and α determines its orientation.

We chose six orientations and three scales for our implementation, obtaining 18 responses $R_{s\alpha}(x, y)$ corresponding to all combinations of $\alpha = 0, 1, \ldots, 5$ and $s = 0, 1, 2$. As in [24], the means $\mu_{s\alpha}$ and standard deviations $\sigma_{s\alpha}$, of the magnitude of the 18 response images were used as features resulting in a descriptor of 36 features.

It is pertinent to mention that σ_x and σ_y from Equation 1 in [24] were computed as $\sigma_x = 1/2\pi\sigma_u$ and $\sigma_y = 1/2\pi\sigma_v$ where σ_u and σ_v are as in Equation 4 in [24]. The scale factor

a ([24], Equation 3) is also computed in accordance with Eq. 4 of [24]. In this equation, the lower and upper frequencies of interest, U_l and U_h, were chosen as 0.05 and 0.4, respectively.

4.3.8 Log-Gabor wavelet features (LGWF)

A variant of the GWF features defined in the previous subsection was derived by substituting log-Gabor wavelet filters [4] for the Gabor wavelet filters. Log-Gabor wavelets are represented in polar coordinates as follows:

$$LG(\rho,\theta) = \exp\left(\frac{\left(-\ln\left(\frac{\rho}{\rho_0}\right)\right)^2}{\left(2\ln\left(\frac{\sigma_\rho}{\rho_0}\right)\right)^2}\right)\cdot\exp\left(\frac{-(\theta-\theta_0)^2}{2\sigma_\theta^2}\right), \tag{7}$$

where ρ is the radial coordinate, θ is the angular coordinate, and ρ_0 is the centre frequency of the filter, defined as $\frac{1}{\lambda}$ for wavelength λ, θ_0 is the orientation of the filter, and σ_ρ and σ_θ are the bandwidths for the radial and angular components.

We used three filter scales $\rho_0 = \{1/3, 1/6, 1/12\}$ corresponding to wavelengths of 3, 6, and 12 pixels and six different orientations $\theta_0 = \{\frac{\pi}{6} * (i-1) \mid i = 1,\ldots,6\}$. For each ρ_0, the radial bandwidth was selected so that $\sigma_\rho/\rho_o = 0.65$, which corresponds approximately to 2 octaves. For all orientations, the angular bandwidth was set to two-third of the angular interval $\pi/6$ (the interval between selected filter orientations θ_0) or $\pi/9$.

The LGWF features were then obtained by computing the mean $\mu_{\rho\theta}$, standard deviation $\sigma_{\rho\theta}$, and skewness $\gamma_{\rho\theta}$ of the magnitude of 18 responses of log-Gabor wavelet filters, defined as follows.

$$\mu_{\rho\theta} = \frac{E(\rho,\theta)}{MN} \tag{8}$$

$$\sigma_{\rho\theta} = \sqrt{\frac{\sum\sum(E(\rho,\theta)-\mu_{\rho\theta})^2}{MN}} \tag{9}$$

$$\gamma_{\rho\theta} = \frac{1}{{\sigma_{\rho\theta}}^3}\cdot\frac{\sum\sum(E(\rho,\theta)-\mu_{\rho\theta})^3}{MN} \tag{10}$$

where M and N are the number of rows and columns of I, and $E(\rho,\theta) = \|\text{idft}(LG(\rho,\theta) * \text{dft}(I))\|$ denotes the magnitude of response image obtained by the pointwise multiplication of discrete Fourier transform of the image I with the log-Gabor wavelet filter. Functions dft and idft denote the forward and inverse discrete Fourier transforms, respectively. The combinations of 3 scales and 6 orientations and 3 features per combination yield a descriptor containing 54 features.

4.3.9 Binary texton features (BTF)

We used the "Binary-MR8" features exactly as described in [16]. These features were chosen since they do not require training. In brief, these features are computed by employing a bank of maximum-response filters consisting of some anisotropic filters at different orientations and scales and some radially symmetric filters. This is followed by a dimensionality reduction process which yields a texture descriptor of length 2048.

4.3.10 Fusion of LGWF and ARPCH (L–A)

LGWF and ARPCH focus on two different aspects of filter responses. The LWGF is based on the global statistics of log-Gabor response image, whereas the ARPCH is based on the local statistics of the phase congruency of the filters response. Various face recognition algorithms focus on the use of phase as well as magnitude-based features of Gabor responses [37] [29]. On the similar lines, considering the complementary (global vs. local) nature of LGWF and ARPCH features, we fused them to be used as a new texture descriptor L–A. Fusion was performed at the "score level" by computing the distance metric (Chi-square) for LGWF and ARPCH separately for each pair of images, normalizing these distances with the min–max normalization described previously, and taking the sum of the normalized distances as the distance between the query image and the database images.

4.3.11 Fusion of LGWF and BTF (L–B)

The filters used in the extraction of binary texton features (BTF) are sensitive to different image features compared to the log-Gabor wavelets used in LGWF. The filter bank used to compute BTF features is rotationally invariant and contains isotropic (Gaussian and Laplacian of Gaussian) as well as anisotropic derivative-based filters which aim to capture edges and bars at multiple orientations and scales and, thus, is well capable of encoding both rotationally invariant and oriented textures. BTF features are derived from the responses of MR8 filters, generating a binary texton for multiple orientations, in a fashion similar to rotation-invariant uniform LBP features. LGWF features are based on response statistics of the log-Gabor filter bank, which consists of log-Gabor filters that are defined in the log-polar coordinates of the Fourier domain as Gaussians shifted from origin, well capable of obtaining the localized frequency information and their zero response for a constant signal provides invariance to greylevel shift. Given the complementary nature of these features, we combined them to create a new texture descriptor L–B for raw material characterization. This fusion is implemented at the score level as described in Sect. 4.3.10.

4.3.12 Fusion of LGWF, ARPCH, and BTF (L–A–B)

We previously observed that LGWF is complementary to ARPCH and BTF. ARPCH and BTF are also complementary. ARPCH is derived from local phase congruency which is an illumination and contrast-invariant feature of an image, whereas BTF is based on anisotropic first and second derivatives, and both of them are based on two different broad approaches to feature analysis in human vision. The possibility of making human feature detection complete by combining insights from both the approaches is listed out in [17]. We fused all three of these descriptors into a new descriptor L–A–B, again using score-level fusion as described in Sect. 4.3.10.

4.3.13 Fusion of LGWF, ARPCH, and SFTA (L–A–S)

Finally, the complementary descriptors LGWF, ARPCH, and SFTA were combined using score-level fusion as in Sect. 4.3.10. This feature is abbreviated as L–A–S.

5 Image-based search algorithm for bifaces

5.1 Feature extraction

The shape descriptor f, the geometric features $g = \{g_1, \ldots, g_7\}$, and all thirteen texture descriptors described in Sect. 4.3 were computed for each image.

5.2 Dissimilarity measures

The distance between each geometric feature of a query image Q and a database image I is defined as a simple absolute difference:

$$D_{\text{geo}}{}^{i}(Q, I) = \left| \left(g_i^Q - g_i^I \right) \right|, \quad i = 1, \ldots, 7, \tag{11}$$

where g_i^Q and g_i^I represent the geometric features of objects in images Q and I, respectively.

The distance between shape descriptors of Q and I is defined as the L_2 norm:

$$D_{\text{shape}}(Q, I) = \sum_{i=1}^{n_{\text{shape}}} \sqrt{\left(f_i^Q - f_i^I \right)^2}, \tag{12}$$

where f_i^Q and f_i^I represent the shape features of images Q and I, respectively, and n_{shape} is the number of shape features, which, as previously noted, is 39.

The distance between texture descriptors of images Q and I is defined as the Chi-square distance:

$$D_{\text{texture}}(Q, I) = \sum_{i=1}^{n_{\text{tex}}} \frac{\left(H_i^Q - H_i^I \right)^2}{H_i^Q + H_i^I}, \tag{13}$$

where H_i^Q and H_i^I represent the features of the texture descriptors of images Q and I, respectively, and n_{tex} is the length of the texture descriptor.

The min–max normalization technique [18] is employed to separately normalize each of the geometric feature distances:

$$D_{\text{geo}}^{\text{norm}(i)}(Q, I) = \frac{D_{\text{geo}}^i(Q, I) - \min_k \left(D_{\text{geo}}^i(Q, I_k) \right)}{\max_k \left(D_{\text{geo}}^i(Q, I_k) \right) - \min_k \left(D_{\text{geo}}^i(Q, I_k) \right)}, \quad i = 1, 2 \ldots, 7 \tag{14}$$

where the index k is over all images in the database. Shape and texture distances, D_{shape} and D_{texture}, are normalized in the same way.

1. Preprocess the query image as in Sect. 4.1.
2. For the input query image, compute the geometric, shape, features described in Sect. 4.2 (these are precomputed for the database images).
3. For the query image, extract the texture features for one of the texture descriptors in Sect. 4.3 (these are precomputed for the database images).
4. Apply Eq. 11 to find $D_{\text{geo}}^i(Q, I)$ for $i = 1, 2, ..., 7$ between query image Q and each database image I. Normalize these distances using min–max normalization (Eq. 14).
5. Apply Eqs. 12 and 13 to find $D_{\text{shape}}(Q, I)$ and $D_{\text{texture}}(Q, I)$ between the query image Q and each database image I. Normalize these distances using min–max normalization.
6. Fuse the 9 normalized distances between Q and each database image I by taking their sum.
7. Based on the fused distances, return the N least dissimilar matches from the database.

Algorithm 1 Proposed algorithm for archaeological image retrieval system.

5.3 Retrieval algorithm

Given a query image, the steps mentioned in Algorithm 1 are performed to retrieve similar images from the database. In order to avoid online computation, the geometric, shape, and texture descriptors of the objects in the database images are precomputed. The algorithm can be used with any one of the texture descriptors described in Sect. 4.3 together with the geometric features g and the shape descriptor f.

6 Validation of the image-based search algorithm

This section discusses the performance evaluation of Algorithm 1.

6.1 Validation methodology

For each texture descriptor, the 1167 front-view images were used in a leave-one-out style methodology. Each of the images was used once as the query image and the N most similar images were retrieved from the set of remaining images using Algorithm 1.

Section 2.1 briefly described the metadata associated with each image. Certain fields of this metadata were used to analyse the performance of Algorithm 1 by comparing the metadata of the retrieved images to the metadata of the query image. Many of the metadata fields contain information that is not useful for image-based matching because they do not directly describe appearance of the objects, for example, fields that detail the date and location of the find. The metadata fields chosen for the use in the analysis were motivated by the findings of the CI study detailed in Sect. 3. The *biface type* (e.g. hand axe) field was used due to its relationship with biface morphology and the blade vs. flake categorization. The *raw material type* field (e.g. flint or chert) was used because this is related to geographic location and availability of material at a site. Several metadata fields describing physical measurements of the artefacts were used since they can be used to judge how well the physical measurements of the bifaces are extracted from the images. These fields were: *Area*, *Aspect Ratio*, *Length*, *Breadth*, *Breadth at 20% of the length* (*Breadth20*, the breadth of the artefact at 20% of the distance between the wider end of the artefact to the narrower end of the biface), *Breadth at 80% of the length (Breadth80)*, and the ratio of *Breadth80* to *Breadth20*.

6.2 Individual metadata match performance

As a first look at how the different texture descriptors performed, for each texture descriptor t and metadata field m, we computed the percentage of retrieved images over all queries where field m of the retrieved image matched field m of the query image. The results are shown in Table 2 for the 10, 20, and 99 most similar images retrieved by Algorithm 1. The biface type field was matched about 93% of the time for all texture descriptors; the lack of variability observed coincides with the expectation that texture descriptors do not capture information about biface type.

Different texture descriptors exhibited greater differences in ability to match raw material type ranging from about 71% for ULBP, OCLBP, and GWF to about 76% for L–A–S for the top 99 retrievals. Generally, the various descriptors consisting of descriptors fused with LWGF resulted in the most raw material type matches, with SFTA and ARPCH also performing well on their own. For 10 and 20 retrievals the same trends are present, though the spread increases to a low of about 72% to a high of about 81% for 10 retrievals. This indicates that the descriptors which are fusions with LWGF are more likely to return bifaces of the same raw material type with higher ranks.

Even though the texture descriptors are not capturing information about a biface's physical dimensions, the trend was that ULBP, and OCLBP resulted in more matches for the physical measurement metadata fields, while the fusions of descriptors with LGWF performed the worst, though in all cases the difference between best and worst was only about four percentage points. This suggests that texture descriptors that are better at matching raw material type do so at the expense of finding bifaces of a more similar size, likely because bifaces of the same raw material type but with greater differences in physical dimensions are selected over bifaces of more similar physical dimensions. The last two columns of the table where we have aggregated the results for metadata features meant to be, respectively, captured by shape and size image features, and we can see the trade-off clearly. Texture descriptors that are better able to extract bifaces with the correct raw material type do so at the expense of matching objects of more similar shape and size.

6.3 Normalized accuracy

In order to get a better picture of the performance of the retrieval system we must recognize that, in general, it will not be possible to get very high percentages of metadata matches since there may not even exist sufficiently many relevant images with similar metadata. In this section we use a measure we call *normalized accuracy* which compares the number of metadata matches in the retrieved images to the highest number of metadata matches that it would be possible to achieve by retrieving the same number of images.

The relevance score, rel_{score}, which characterizes the relevance of an image D retrieved by Algorithm 1 to the given query image Q is defined as:

$$rel_{\text{score}}(Q, D) = \sum_{i=1}^{9} meta^{i}_{\text{score}}(Q, D), \tag{15}$$

where $meta^{i}_{\text{score}}(Q, D)$ is the binary match score function

$$meta^{i}_{\text{score}}(Q, D) = \begin{cases} 1, & \text{if } meta^{i}(Q) = meta^{i}(D), \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

where $meta^{i}(Q)$ and $meta^{i}(D)$ refer to the value of the i^{th} metadata feature of Q and D, respectively. Thus, if the two images exactly match in all of their metadata features, the

Table 2 Percentage of queries for which each metadata field matched for each texture descriptor

Top 99 Retreivals											
	Shape			Size					Texture		
Descriptor	Type	Breadth Ratio	Aspect	Area	Breadth20	Breadth80	Length	Breadth	Raw Mat.	All Shape	All Size
ULBP	93.66	53.92	55.1	70.61	50.44	62.54	66.95	64.87	71.13	67.56	63.08
OCLBP	93.66	53.98	55.21	70.66	50.52	62.56	67.08	64.94	71.2	67.62	63.15
SFTA	93.65	53.62	51.58	70.4	49.92	62.09	65.11	64.34	74.85	66.28	62.37
GPCH	93.69	53.99	53.52	70.75	50.25	62.61	66.35	65.05	72.65	67.07	63.0
ARPCH	93.73	54.13	52.5	70.38	50.3	62.15	65.61	64.55	73.05	66.79	62.6
OPCH	93.63	53.23	54.93	70.12	49.98	62.04	66.6	64.39	72.31	67.26	62.63
GWF	93.57	53.32	52.95	70.11	49.56	62.0	65.88	64.5	70.83	66.61	62.41
LGWF	93.57	52.48	52.01	69.2	48.72	61.17	64.89	63.45	71.67	66.02	61.49
BTF	93.66	53.18	56.21	70.13	49.72	61.69	66.96	63.88	71.89	67.68	62.48
L-A	93.65	51.9	51.14	68.01	47.73	59.68	63.74	61.9	73.4	65.56	60.21
L-B	93.58	51.34	54.93	68.0	47.55	59.53	65.12	61.55	72.33	66.62	60.35
L-A-B	93.66	50.83	54.08	66.75	46.71	58.3	63.82	60.13	73.95	66.19	59.14
L-A-S	93.61	51.02	50.21	66.94	46.71	58.68	62.47	60.38	76.2	64.95	59.03

Top 20 Retreivals											
	Shape			Size					Texture		
Descriptor	Type	Breadth Ratio	Aspect	Area	Breadth20	Breadth80	Length	Breadth	Raw Mat.	All Shape	All Size
ULBP	93.54	67.77	70.37	80.86	64.91	72.77	79.01	75.57	71.54	77.22	74.63
OCLBP	93.59	67.87	70.65	80.86	65.02	72.73	79.1	75.64	71.59	77.37	74.67
SFTA	93.57	66.18	66.66	80.36	63.37	72.14	76.99	74.53	76.91	75.47	73.48
GPCH	93.56	67.62	69.19	80.5	64.19	72.7	78.33	75.68	74.3	76.79	74.28
ARPCH	93.56	67.06	67.81	80.07	63.99	72.01	77.32	74.95	75.02	76.15	73.67
OPCH	93.53	67.16	69.83	80.46	64.59	72.55	78.65	75.32	73.44	76.84	74.31
GWF	93.35	66.91	67.79	79.88	63.72	72.41	77.88	74.96	71.38	76.02	73.77
LGWF	93.44	65.49	66.27	78.93	62.31	71.52	76.99	74.15	73.42	75.07	72.78
BTF	93.62	66.97	71.02	80.24	64.26	72.52	78.92	75.07	73.33	77.2	74.2
L-A	93.48	64.22	64.72	77.99	61.24	70.3	75.38	73.09	76.5	74.14	71.6
L-B	93.53	63.99	68.27	78.12	61.34	70.45	77.0	73.01	75.0	75.26	71.98
L-A-B	93.56	62.87	66.95	77.27	60.39	69.32	75.96	71.93	77.73	74.46	70.97
L-A-S	93.55	62.7	63.57	77.24	59.72	69.14	74.21	71.3	79.75	73.27	70.32

Top 10 Retreivals											
	Shape			Size					Texture		
Descriptor	Type	Breadth Ratio	Aspect	Area	Breadth20	Breadth80	Length	Breadth	Raw Mat.	All Shape	All Size
ULBP	93.32	72.35	74.82	83.16	69.59	75.64	81.97	78.76	71.87	80.16	77.82
OCLBP	93.33	72.54	74.87	83.4	69.61	75.78	82.21	78.91	71.71	80.25	77.98
SFTA	93.59	70.59	71.95	82.95	67.27	75.47	80.18	77.69	77.13	78.71	76.71
GPCH	93.54	72.38	73.54	82.94	68.89	75.52	81.08	78.4	75.23	79.82	77.37
ARPCH	93.51	71.6	72.21	82.58	68.71	75.15	80.47	77.69	75.85	79.11	76.92
OPCH	93.56	71.89	74.18	82.86	69.48	75.35	81.74	78.56	73.63	79.88	77.6
GWF	93.23	71.98	72.55	82.3	68.78	75.51	80.46	78.06	71.6	79.25	77.02
LGWF	93.28	69.9	70.89	81.55	66.7	74.8	79.6	77.21	74.26	78.02	75.97
BTF	93.61	71.45	75.21	82.91	69.02	75.52	82.02	78.83	74.15	80.09	77.66
L-A	93.32	68.76	69.36	80.6	66.13	73.49	78.74	76.08	77.55	77.15	75.01
L-B	93.37	68.9	72.0	80.69	65.87	74.04	79.83	76.42	76.12	78.09	75.37
L-A-B	93.47	67.5	70.61	79.89	65.09	72.62	78.69	75.21	78.96	77.2	74.3
L-A-S	93.44	67.19	67.69	80.24	63.94	72.42	77.51	74.76	81.07	76.11	73.77

The last two columns group the results for metadata meant to be captured by size and shape features, respectively

relevance score would be 9 and the image D is considered highly relevant to the query image Q.

The maximum possible total relevance score, denoted as $max_{\text{score}}(Q)$, for a query image Q and N retrievals is determined by computing $rel_{\text{score}}(Q, D)$ for every database image D and finding images $E_1, \ldots, E_N$ with the N largest relevance scores. Then $max_{\text{score}}(Q) = \sum_{i=1}^{N} rel_{\text{score}}(Q, E_i)$. Similarly, the actual total relevance score achieved by Algorithm 1 for query image Q is $query_{\text{score}}(Q) = \sum_{i=1}^{N} rel_{\text{score}}(Q, D_i)$, where $D_1, \ldots, D_N$ are the N images that were retrieved. Finally, the *normalized accuracy* for a given query Q is:

$$Ret_{\text{accuracy}}(Q) = \frac{query_{\text{score}}(Q)}{max_{\text{score}}(Q)}. \tag{17}$$

For each texture descriptor, the normalized accuracy was computed for each query using the leave-one-out methodology described in Sect. 6.1 at $N = 99$ retrievals. The mean normalized accuracy for each texture descriptor over all queries is shown in Fig. 4. A Wilcoxon signed-rank test was performed pairwise on the texture descriptors to test the null hypothesis that the 1167 paired normalized accuracy scores came from the same distribution. In nearly all cases the null hypothesis was rejected at the $\alpha = 0.05$ level with $p < 0.0069$ with the exceptions of ULBP/GPCH ($p = 0.0733$) and OPCH/BTF ($p = 0.9236$). Results for $N = 10$ and $N = 20$ retrievals were extremely similar, both in the magnitude of the means for each descriptor and the results of the Wilcoxon signed-rank tests. These results indicate that all of the texture descriptors (along with the shape and geometric features) are doing very well overall in matching the most relevant images. The differences in performance, while small in magnitude, are statistically significant and indicate that ULBP, OCLBP, GPCH, and BTF are the best texture descriptors on average.

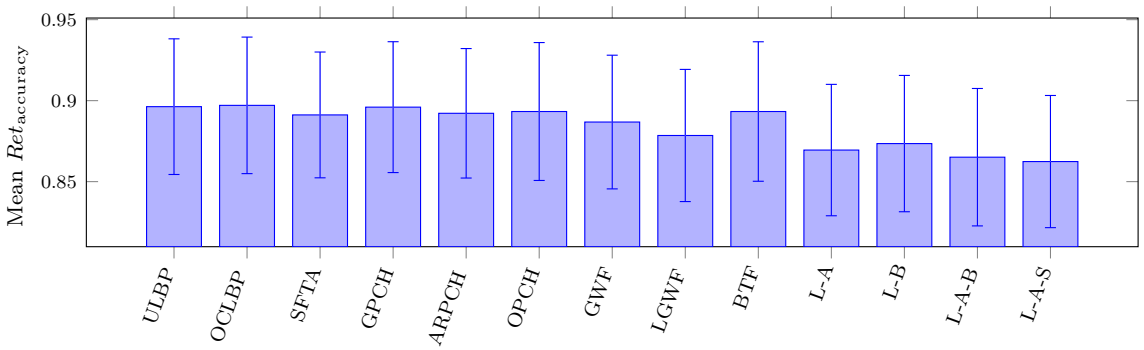


Fig. 4 Mean and standard deviation of normalized accuracy for each texture descriptor over all queries using Algorithm 1, $N = 99$ retrievals

However, given that this image retrieval system is intended to support user search and reasoning over archives, looking only at differences in mean performance does not provide a sufficiently nuanced understanding of the differences between the algorithms and what the consequences of those differences might be for the search experience. In the next section, we propose an alternative analysis that examines the data sets from a differing perspective.

For the remainder of our analysis we will be interested in the behaviour of the ULBP, OLCBP, GPCH, and BTF as the group of descriptors that had the "best" overall performance in Table 4 which will refer to as the SO group (superior overall group) and the fused descriptors L–A, L–B, L–A–B, and L–A–S as the IO group (inferior overall group) which had the "worst" performance in Fig. 4.

6.4 Pairwise symmetric difference comparisons

All of the texture descriptors resulted in similar overall accuracy (Fig. 4). When the overall accuracy score is within two to three percentage points, it is tempting to say that any particular algorithm would be suitable. However, because this algorithm is intended to support user search, these aggregate measures of accuracy do not give us any indication of what result sets look like which could be very important in choosing a particular algorithm. There could be distinct differences in the sets of images that are returned because when comparing any two descriptors across the nearly 1200 query images, there are a number of different distributions of results that could produce similar overall accuracy scores. If there are situations where one descriptor performs particularly well or particularly poorly this may skew the average up or down. Similarly, if the result sets retrieved by two different descriptors consistently share a number of results, this will obscure the actual differences between the algorithms.

To better elucidate the previously observed trade-offs apparent in the set of texture descriptors we performed an analysis in which, for a given query image Q and pair of texture descriptors D_1 and D_2, we only considered the symmetric difference of the images retrieved by Algorithm 1 using D_1 and D_2, that is, images not in common to both retrievals.

For a pair of descriptors D_1 and D_2, let R_q^1 and R_q^2 be the set of images retrieved for a given query image q by Algorithm 1 using D_1 and D_2, respectively. Obtain the symmetric difference of R_q^1 and R_q^2: $\Delta_q(D_1, D_2) = R_q^1 \cup R_q^2 \setminus R_q^1 \cap R_q^2$. Let $\delta_q^1(D_1, D_2)$ be the elements of $\Delta_q(D_1, D_2)$ in R_1^1, and let $\delta_q^2(D_1, D_2)$ be the elements of $\Delta_q(D_1, D_2)$ in R_q^2.

6.4.1 Average number of disagreements

If two sets of results for a given query image using different descriptors are identical, then they are performing nearly the same for that image. For retrieved sets that are small (e.g. 10 retrievals), then it would be expected that most results would be shared. If two descriptors return largely the same set of images, then we can assume there is no major difference between them and that the descriptors are doing essentially the same thing. For larger retrieval sets, there are fewer appropriate matches that will occur, so it is valuable to examine which feature performs best after removing shared items.

Table 3 presents the results for each pair of descriptors for the top $N = 10$ and $N = 99$ retrievals. Looking closely at the results for the SO feature group, we see that these descriptors are all performing nearly identically in that the largest average difference between retrieval sets for pairs of descriptors in this group is 2.46 (GPCH vs. BTF). However, when comparing these 4 descriptors to those in the IO group, we find that there are substantial differences between the retrieval sets for the IO group and the SO group with ULBP/L–A–S producing the highest mean difference of 5.2 images between the sets.

When we increase to 99 retrievals we find similarities between the IO group and the SO group that are even more pronounced. Pairwise, all of ULBP, OCLBP, and GPCH produce result sets that have fewer than 13 different images on average. As a result, it is hardly worth calculating the dif-

Table 3 Average size of the symmetric difference for the top 10 and top 100 retrievals

Top 10 Retreivals – Avg. Number of Disagreements

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	0.69	3.18	1.63	2.42	1.60	2.49	3.54	1.82	4.17	3.76	4.34	4.98
OCLBP	0.69	0.00	3.10	1.88	2.70	1.59	2.63	3.76	1.68	4.45	3.99	4.61	5.20
SFTA	3.18	3.10	0.00	3.20	3.60	3.29	3.70	4.15	3.11	4.64	4.21	4.71	4.15
GPCH	1.63	1.88	3.20	0.00	1.61	2.14	2.69	3.34	2.46	3.55	3.61	3.84	4.43
ARPCH	2.42	2.70	3.60	1.61	0.00	2.77	3.04	3.42	3.04	3.05	3.65	3.39	4.03
OPCH	1.60	1.59	3.29	2.14	2.77	0.00	3.02	3.86	2.10	4.44	4.03	4.54	5.13
GWF	2.49	2.63	3.70	2.69	3.04	3.02	0.00	2.32	3.04	3.33	2.79	3.63	4.25
LGWF	3.54	3.76	4.15	3.34	3.42	3.86	2.32	0.00	3.87	1.92	1.51	2.41	3.13
BTF	1.82	1.68	3.11	2.46	3.04	2.10	3.04	3.87	0.00	4.49	3.57	4.26	5.08
L-A	4.17	4.45	4.64	3.55	3.05	4.44	3.33	1.92	4.49	0.00	2.24	1.40	2.21
L-B	3.76	3.99	4.21	3.61	3.65	4.03	2.79	1.51	3.57	2.24	0.00	1.75	3.09
L-A-B	4.34	4.61	4.71	3.84	3.39	4.54	3.63	2.41	4.26	1.40	1.75	0.00	2.34
L-A-S	4.98	5.20	4.15	4.43	4.03	5.13	4.25	3.13	5.08	2.21	3.09	2.34	0.00

0 1.7 3.3 5

Top 99 Retreivals – Avg. Number of Disagreements

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	2.65	17.63	8.92	12.98	9.89	14.16	20.12	7.88	25.05	20.52	25.05	30.30
OCLBP	2.65	0.00	17.65	9.92	14.27	9.83	14.93	21.20	7.29	26.44	21.56	26.40	31.53
SFTA	17.63	17.65	0.00	17.35	19.52	19.68	20.79	23.83	18.13	27.62	24.66	28.05	24.62
GPCH	8.92	9.92	17.35	0.00	8.23	14.36	14.38	18.73	12.92	21.23	20.44	22.51	26.62
ARPCH	12.98	14.27	19.52	8.23	0.00	17.03	16.67	19.45	16.17	18.59	21.07	20.29	24.48
OPCH	9.89	9.83	19.68	14.36	17.03	0.00	18.46	23.69	11.73	28.24	23.74	27.92	32.76
GWF	14.16	14.93	20.79	14.38	16.67	18.46	0.00	11.76	16.69	18.66	14.77	20.20	24.79
LGWF	20.12	21.20	23.83	18.73	19.45	23.69	11.76	0.00	22.01	11.02	8.97	13.85	18.28
BTF	7.88	7.29	18.13	12.92	16.17	11.73	16.69	22.01	0.00	26.83	19.70	24.82	31.07
L-A	25.05	26.44	27.62	21.23	18.59	28.24	18.66	11.02	26.83	0.00	13.44	8.48	12.82
L-B	20.52	21.56	24.66	20.44	21.07	23.74	14.77	8.97	19.70	13.44	0.00	10.04	18.70
L-A-B	25.05	26.40	28.05	22.51	20.29	27.92	20.20	13.85	24.82	8.48	10.04	0.00	14.09
L-A-S	30.30	31.53	24.62	26.62	24.48	32.76	24.79	18.28	31.07	12.82	18.70	14.09	0.00

0 20 40 60

ferences in average accuracy in these algorithms as they are doing largely the same thing.

Comparing the SO group to the IO group we see a much more pronounced difference. OCLBP and L–A–S have 31.5 different images (nearly a third) in their result sets on average. Most pairs consisting of one IO and one SO group member average between 10 and 25 disagreements (10.1–25.3% of retrievals). These results show that on average, there are a non-trivial number of differences between the set of results returned by different descriptors, but these are masked in the overall accuracy result. Given these differences, it is worth looking closely at the differences, first by looking at the mean accuracy of the differences.

6.4.2 Differences in metadata matches for symmetric differences

We computed the difference in the percentage of metadata matches for all of the images in the symmetric differences of each query. The process is now explained in detail.

Let $m_q^1(D_1, D_2)$ be the number of metadata matches between the elements of R_q^1 and the query image q, and $n_q^1(D_1, D_2) = 9|R_q^1|$ (the total number of metadata fields associated with the elements of R_q^1). Symmetrically define $m_q^2(D_1, D_2)$ and $n_q^2 = 9|R_q^2|$. Then compute:

$$P(D_1, D_2) = \frac{\sum_q m_q^1(D_1, D_2)}{\sum_q n_q^1(D_1, D_2)} - \frac{\sum_q m_q^2(D_1, D_2)}{\sum_q n_q^2(D_1, D_2)} \quad (18)$$

which is the difference between the percentage of metadata matches in the symmetric difference of query results arising from descriptors D_1 and D_2. The two subtractive terms in Eq. 18 are relevance scores for D_1 and D_2, and we call $P(D_1, D_2)$ the *relevance score difference*.

Table 4a shows the relevance score difference for all pairs of texture descriptors for queries using Algorithm 1. The rows are indexed by D_1 and columns by D_2; thus, the top-right most entry indicates that the percentage of metadata fields matched was 5% more for ULBP than for L–A–S. This table confirms the observations from Table 2 that the descriptors in the SO group have the best overall performance (most entries in their rows are positive).

Table 4b shows the same relevance score difference but only considering the single metadata field *Breadth at 20% Length* (definitions of m_q^1, m_q^2, n_q^1, and n_q^2 are adjusted accordingly) rather than all metadata fields together. The results indicate that the SO group are generally overall strong performers for matching this metadata field. Results are similar for the other geometric features in g.

Table 4c shows the relevance score difference considering only the metadata field *biface type*. The data indicate that the choice of texture descriptor has very little effect on the

Table 4 Relevance score difference comparisons for top 10 retrievals

(a) Top 10 Retreivals – Relevance Score Difference – All Metadata

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	-1.44	1.63	-0.02	1.70	0.16	3.13	4.17	-0.76	4.65	4.21	4.98	5.18
OCLBP	1.44	0.00	1.99	0.51	1.89	0.79	3.35	4.19	-0.23	4.58	4.22	4.90	5.15
SFTA	-1.63	-1.99	0.00	-1.63	-0.29	-1.50	0.71	2.32	-2.11	3.07	2.53	3.49	4.97
GPCH	0.02	-0.51	1.63	0.00	2.59	0.14	2.91	4.44	-0.55	5.48	4.39	5.64	5.84
ARPCH	-1.70	-1.89	0.29	-2.59	0.00	-1.40	1.21	3.12	-1.81	5.00	3.21	5.15	5.39
OPCH	-0.16	-0.79	1.50	-0.14	1.40	0.00	2.50	3.76	-0.78	4.31	3.86	4.70	4.98
GWF	-3.13	-3.35	-0.71	-2.91	-1.21	-2.50	0.00	3.01	-3.02	3.49	2.88	3.80	4.24
LGWF	-4.17	-4.19	-2.32	-4.44	-3.12	-3.76	-3.01	0.00	-4.17	2.41	0.70	2.83	3.53
BTF	0.76	0.23	2.11	0.55	1.81	0.78	3.02	4.17	0.00	4.62	4.83	5.40	5.35
L-A	-4.65	-4.58	-3.07	-5.48	-5.00	-4.31	-3.49	-2.41	-4.62	0.00	-1.59	1.58	2.90
L-B	-4.21	-4.22	-2.53	-4.39	-3.21	-3.86	-2.88	-0.70	-4.83	1.59	0.00	3.30	3.24
L-A-B	-4.98	-4.90	-3.49	-5.64	-5.15	-4.70	-3.80	-2.83	-5.40	-1.58	-3.30	0.00	1.80
L-A-S	-5.18	-5.15	-4.97	-5.84	-5.39	-4.98	-4.24	-3.53	-5.35	-2.90	-3.24	-1.80	0.00

(b) Top 10 Retreivals – Relevance Score Difference – Breadth at 20% Length

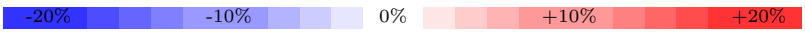
	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	-0.38	7.28	4.31	3.62	0.66	3.24	8.17	3.13	8.30	9.88	10.37	11.34
OCLBP	0.38	0.00	7.54	3.87	3.34	0.83	3.17	7.76	3.55	7.84	9.39	9.81	10.90
SFTA	-7.28	-7.54	0.00	-5.04	-3.99	-6.71	-4.08	1.39	-5.61	2.48	3.33	4.64	8.03
GPCH	-4.31	-3.87	5.04	0.00	1.09	-2.79	0.39	6.57	-0.54	7.79	8.35	9.89	11.17
ARPCH	-3.62	-3.34	3.99	-1.09	0.00	-2.79	-0.23	5.90	-1.01	8.46	7.78	10.66	11.84
OPCH	-0.66	-0.83	6.71	2.79	2.79	0.00	2.32	7.22	2.21	7.56	8.95	9.67	10.79
GWF	-3.24	-3.17	4.08	-0.39	0.23	-2.32	0.00	9.00	-0.78	7.98	10.43	10.17	11.38
LGWF	-8.17	-7.76	-1.39	-6.57	-5.90	-7.22	-9.00	0.00	-6.00	2.97	5.47	6.65	8.80
BTF	-3.13	-3.55	5.61	0.54	1.01	-2.21	0.78	6.00	0.00	6.44	8.82	9.22	9.98
L-A	-8.30	-7.84	-2.48	-7.79	-8.46	-7.56	-7.98	-2.97	-6.44	0.00	1.13	7.40	9.85
L-B	-9.88	-9.39	-3.33	-8.35	-7.78	-8.95	-10.43	-5.47	-8.82	-1.13	0.00	4.46	6.25
L-A-B	-10.37	-9.81	-4.64	-9.89	-10.66	-9.67	-10.17	-6.65	-9.22	-7.40	-4.46	0.00	4.91
L-A-S	-11.34	-10.90	-8.03	-11.17	-11.84	-10.79	-11.38	-8.80	-9.98	-9.85	-6.25	-4.91	0.00

(c) Top 10 Retreivals – Relevance Score Difference – Biface Type

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	-0.13	-0.86	-1.35	-0.80	-1.47	0.39	0.12	-1.59	0.00	-0.12	-0.34	-0.25
OCLBP	0.13	0.00	-0.85	-1.12	-0.68	-1.44	0.40	0.14	-1.67	0.02	-0.09	-0.30	-0.22
SFTA	0.86	0.85	0.00	0.16	0.22	0.11	1.00	0.76	-0.06	0.59	0.54	0.26	0.36
GPCH	1.35	1.12	-0.16	0.00	0.16	-0.08	1.17	0.79	-0.29	0.62	0.49	0.18	0.22
ARPCH	0.80	0.68	-0.22	-0.16	0.00	-0.16	0.95	0.69	-0.32	0.63	0.41	0.13	0.17
OPCH	1.47	1.44	-0.11	0.08	0.16	0.00	1.10	0.73	-0.25	0.53	0.48	0.19	0.22
GWF	-0.39	-0.40	-1.00	-1.17	-0.95	-1.10	0.00	-0.23	-1.27	-0.29	-0.50	-0.68	-0.52
LGWF	-0.12	-0.14	-0.76	-0.79	-0.69	-0.73	0.23	0.00	-0.86	-0.23	-0.58	-0.80	-0.53
BTF	1.59	1.67	0.06	0.29	0.32	0.25	1.27	0.86	0.00	0.64	0.69	0.33	0.33
L-A	-0.00	-0.02	-0.59	-0.62	-0.63	-0.53	0.29	0.23	-0.64	0.00	-0.20	-1.07	-0.55
L-B	0.12	0.09	-0.54	-0.49	-0.41	-0.48	0.50	0.58	-0.69	0.20	0.00	-0.60	-0.26
L-A-B	0.34	0.30	-0.26	-0.18	-0.13	-0.19	0.68	0.80	-0.33	1.07	0.60	0.00	0.11
L-A-S	0.25	0.22	-0.36	-0.22	-0.17	-0.22	0.52	0.53	-0.33	0.55	0.26	-0.11	0.00

(d) Top 10 Retreivals – Relevance Score Difference – Raw Material Type

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	2.29	-16.58	-20.68	-16.48	-10.98	1.09	-6.76	-12.58	-13.63	-11.30	-16.35	-18.49
OCLBP	-2.29	0.00	-17.48	-18.75	-15.33	-12.09	0.43	-6.78	-14.58	-13.15	-11.06	-15.73	-17.99
SFTA	16.58	17.48	0.00	5.95	3.55	10.66	14.98	6.93	9.59	-0.91	2.42	-3.87	-9.48
GPCH	20.68	18.75	-5.95	0.00	-3.88	7.51	13.53	2.92	4.39	-6.55	-2.45	-9.71	-13.19
ARPCH	16.48	15.33	-3.55	3.88	0.00	8.04	14.01	4.67	5.60	-5.57	-0.72	-9.14	-12.95
OPCH	10.98	12.09	-10.66	-7.51	-8.04	0.00	6.73	-1.63	-2.50	-8.85	-6.17	-11.73	-14.50
GWF	-1.09	-0.43	-14.98	-13.53	-14.01	-6.73	0.00	-11.49	-8.43	-17.91	-16.21	-20.28	-22.28
LGWF	6.76	6.78	-6.93	-2.92	-4.67	1.63	11.49	0.00	0.27	-17.18	-12.33	-19.49	-21.77
BTF	12.58	14.58	-9.59	-4.39	-5.60	2.50	8.43	-0.27	0.00	-7.57	-5.51	-11.28	-13.60
L-A	13.63	13.15	0.91	6.55	5.57	8.85	17.91	17.18	7.57	0.00	6.41	-10.03	-15.87
L-B	11.30	11.06	-2.42	2.45	0.72	6.17	16.21	12.33	5.51	-6.41	0.00	-16.22	-16.05
L-A-B	16.35	15.73	3.87	9.71	9.14	11.73	20.28	19.49	11.28	10.03	16.22	0.00	-9.04
L-A-S	18.49	17.99	9.48	13.19	12.95	14.50	22.28	21.77	13.60	15.87	16.05	9.04	0.00



matching of *biface type* which is consistent with the intuition that texture should not be indicative of the type of biface since this is primarily determined by shape.

Table 4d shows the relevance score difference considering only the metadata field *raw material type*. Here we see a dramatic difference in performance where the table colouring resembles an inversion of the other tables. The best performing descriptors in the other tables are among the worst performers for matching raw material type. The descriptors in the IO group are the best performers which is consistent with observations from Table 2.

This set of results demonstrates that the descriptors in the IO group are better at matching *biface type* at the expense of other metadata fields. The more pronounced differences in performance between descriptors observed in all of these tables, relative to Table 2, are because we are considering

only images in the symmetric differences of the retrievals, that is, the query results for which two different descriptors disagree.

For the top 99 retrievals the results are similar to those for the top 10 retrievals.

6.4.3 Dominance

As we can see from the above results, there are distinct differences in the mean accuracy of the symmetric difference. However, like the overall accuracy calculations, these averages may be hiding nuances about the symmetric difference sets. If, for example, one descriptor consistently performs better in its results, then it could be argued to choose one over another. As a result, we considered the question of how often does one descriptor perform better than another descriptor which we will refer to as *dominance* which is calculated using the A-primed statistic [35].

For query image q and descriptors D_1 and D_2, we rank the elements of the symmetric difference of the query results $\Delta_q(D_1, D_2)$ according to the number of metadata matches with the query image as they would be in a Mann–Whitney test. The image in $\Delta_q(D_1, D_2)$ with the greatest number of metadata matches gets rank 1, the next most gets rank 2, and so on. In the case of ties, the tied images get the average of the rank they occupy, so if three images that occupy ranks 7, 6, 5, and 4 have the same number of metadata matches, they are all assigned rank 5.5. We then calculate the sum of these ranks for the images from $\Delta_q(D_1, D_2)$ that are in R_q^1, which we previously named $\delta_q^1(D_1, D_2)$ (see Sect. 6.4) and denote this as *ranksum*. Then, *dominance* for query q is:

$$A_q(D_1, D_2) = \frac{\left(\frac{ranksum}{|\delta_q^1(D_1, D_2)|} - \left(|\delta_q^1(D_1, D_2) + 1\right)/2\right)}{|\delta_q^2(D_1, D_2)|} \tag{19}$$

The results for dominance are shown in Table 5 for the top $N = 10$ retrievals, respectively. The top-most subtable shows the average dominance (on average how often does one descriptor perform better than another), and the middle subtable shows the percentage of queries where one descriptor performs better than another. The bottom-most subtable shows how often one descriptor has a "big win" over another descriptor, by which we mean the percentage of queries for which a descriptor has a dominance $\geq$0.8 over the descriptor to which it is being compared.

From Table 5 we can see that when features in the SO group are compared to each other, the dominance scores are are all approximately 50%. For these descriptors, if you choose two at random you have about an equal chance of getting the most accurate return set for a given query image.

For the IO group there is a gradual progression of L–A–S dominating the other three, with the smallest difference being between L–A–B and L–A–S. When we compare descriptors in the IO group with those in the SO group, we find that those in the SO group dominate approximately 60% of the time. However, considering "big wins" (Table 5b), we find that the SO group query results rarely dominate the IO group query results with a dominance of more than 0.8. That is, while the SO group on average produces results with more metadata matches than the IO group, it is rare that the SO consistently produces many more metadata matches. This tends to indicate that the IO descriptors are returning a comparable set of results to the SO group results but with a different set of images that are less focused on matching the metadata precisely.

6.5 Discussion and conclusions

Our results have built a robust picture of the strengths and weaknesses of the different descriptors. Considering overall accuracy, it does appear that the descriptors perform largely the same in the way that they match images, with less than a percentage point of accuracy between the top 5 methods. This means that if we were worried strictly about finding as close to perfect matches as we can, then any of these descriptors would likely be adequate. This is particularly true when there are small numbers of images in the sets that have exact matches against the ground truth metadata. For example, if there are only 7 objects that perfectly match in the ground truth to the query image, we can be relatively certain each of the algorithms will return those 7 reliably.

However, given that we are looking to support online search, we are interested in more than just exact matches. Researchers working with image archives rely on comparing and contrasting different, yet related, objects together while making their decisions. As a result, it was important to understand whether each descriptor was just returning the same sets as every other descriptor or whether there were distinct differences in their performance. There was distinct difference between the best and worst overall performing descriptors (the SO group and the IO group). First, it is interesting that in the SO group there is almost no variation in the sets of images they retrieve. Even at their most varied, we are only seeing about 13% variation in the images that are returned from the standard descriptors, with most hovering around only 8% difference. This tends to imply that if you choose any one of these algorithms, you are going to get largely the same set of images being returned.

In comparison, the IO group of descriptors retrieves image sets that are 20–30% different from the SO group. It appears that the IO descriptors are better at detecting differences in material while trading off shape matching accuracy, whereas the SO descriptors are doing the opposite. If we are working

Table 5 Average dominance, frequency of dominance, and frequency of "big wins" for top 10 retrievals

Top 10 Retreivals - Avg. Dominance over all queries (closer to 1.0 means descriptor in the **row** dominates)

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	0.48	0.54	0.51	0.54	0.49	0.55	0.58	0.49	0.60	0.58	0.61	0.61
OCLBP	0.52	0.00	0.54	0.52	0.54	0.51	0.56	0.58	0.50	0.60	0.58	0.61	0.61
SFTA	0.46	0.46	0.00	0.46	0.49	0.46	0.50	0.54	0.45	0.56	0.55	0.57	0.60
GPCH	0.49	0.48	0.54	0.00	0.56	0.49	0.55	0.59	0.49	0.62	0.59	0.62	0.63
ARPCH	0.46	0.46	0.51	0.44	0.00	0.46	0.51	0.57	0.46	0.61	0.57	0.61	0.62
OPCH	0.51	0.49	0.54	0.51	0.54	0.00	0.55	0.58	0.49	0.59	0.58	0.60	0.61
GWF	0.45	0.44	0.50	0.45	0.49	0.45	0.00	0.57	0.43	0.58	0.57	0.59	0.60
LGWF	0.42	0.42	0.46	0.41	0.43	0.42	0.43	0.00	0.41	0.55	0.51	0.56	0.58
BTF	0.51	0.50	0.55	0.51	0.54	0.51	0.57	0.59	0.00	0.60	0.60	0.62	0.62
L-A	0.40	0.40	0.44	0.38	0.39	0.41	0.42	0.45	0.40	0.00	0.47	0.53	0.57
L-B	0.42	0.42	0.45	0.41	0.43	0.42	0.43	0.49	0.40	0.53	0.00	0.57	0.58
L-A-B	0.39	0.39	0.43	0.38	0.39	0.40	0.41	0.44	0.38	0.47	0.43	0.00	0.54
L-A-S	0.39	0.39	0.40	0.37	0.38	0.39	0.40	0.42	0.38	0.43	0.42	0.46	0.00

Top 10 Retreivals - % of Queries for which descriptor in the **row** dominates.

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	0.18	0.48	0.34	0.47	0.32	0.48	0.62	0.33	0.66	0.61	0.66	0.70
OCLBP	0.23	0.00	0.49	0.37	0.50	0.33	0.49	0.63	0.34	0.67	0.63	0.68	0.71
SFTA	0.36	0.36	0.00	0.37	0.42	0.37	0.45	0.55	0.35	0.60	0.57	0.62	0.66
GPCH	0.32	0.31	0.48	0.00	0.42	0.37	0.49	0.59	0.37	0.65	0.61	0.67	0.72
ARPCH	0.33	0.33	0.43	0.26	0.00	0.35	0.43	0.55	0.34	0.61	0.57	0.63	0.67
OPCH	0.35	0.31	0.49	0.39	0.47	0.00	0.50	0.61	0.36	0.65	0.61	0.67	0.69
GWF	0.31	0.30	0.42	0.32	0.41	0.33	0.00	0.52	0.30	0.57	0.53	0.60	0.65
LGWF	0.27	0.27	0.35	0.28	0.32	0.28	0.29	0.00	0.27	0.42	0.35	0.49	0.56
BTF	0.38	0.34	0.50	0.42	0.49	0.38	0.52	0.63	0.00	0.67	0.63	0.69	0.72
L-A	0.26	0.26	0.32	0.25	0.26	0.27	0.30	0.30	0.27	0.00	0.34	0.35	0.49
L-B	0.28	0.27	0.34	0.28	0.33	0.29	0.31	0.32	0.26	0.44	0.00	0.44	0.54
L-A-B	0.25	0.25	0.31	0.24	0.27	0.25	0.28	0.31	0.23	0.30	0.27	0.00	0.44
L-A-S	0.25	0.24	0.25	0.22	0.24	0.24	0.28	0.30	0.22	0.29	0.29	0.34	0.00

Top 10 Retreivals - % of Queries for which descriptor in the **row** has dominance ≥ 0.8).

	ULBP	OCLBP	SFTA	GPCH	ARPCH	OPCH	GWF	LGWF	BTF	L-A	L-B	L-A-B	L-A-S
ULBP	0.00	0.15	0.17	0.20	0.22	0.19	0.21	0.19	0.18	0.19	0.18	0.20	0.18
OCLBP	0.18	0.00	0.18	0.20	0.19	0.20	0.22	0.18	0.20	0.19	0.18	0.19	0.17
SFTA	0.11	0.11	0.00	0.10	0.10	0.11	0.11	0.12	0.11	0.13	0.12	0.14	0.20
GPCH	0.19	0.18	0.17	0.00	0.27	0.18	0.21	0.22	0.16	0.25	0.21	0.25	0.21
ARPCH	0.14	0.14	0.12	0.17	0.00	0.13	0.16	0.19	0.11	0.25	0.18	0.26	0.21
OPCH	0.21	0.20	0.18	0.19	0.19	0.00	0.18	0.18	0.17	0.17	0.18	0.19	0.16
GWF	0.14	0.13	0.12	0.14	0.14	0.11	0.00	0.24	0.09	0.20	0.21	0.20	0.20
LGWF	0.07	0.07	0.07	0.07	0.09	0.06	0.12	0.00	0.05	0.24	0.22	0.24	0.22
BTF	0.21	0.21	0.18	0.17	0.17	0.20	0.20	0.18	0.00	0.17	0.21	0.20	0.17
L-A	0.04	0.04	0.05	0.06	0.08	0.04	0.08	0.16	0.04	0.00	0.16	0.26	0.25
L-B	0.06	0.07	0.08	0.07	0.08	0.06	0.10	0.19	0.06	0.21	0.00	0.28	0.21
L-A-B	0.05	0.04	0.04	0.06	0.07	0.05	0.06	0.12	0.04	0.19	0.16	0.00	0.21
L-A-S	0.03	0.03	0.05	0.03	0.04	0.03	0.04	0.08	0.03	0.13	0.08	0.15	0.00

0 .21 .42 .63

with a single collection of flint, where most pieces come from the same source stone, then the SO descriptors are likely to perform better. However, for a more varied collection of bifaces, comprised of flint, glass, etc., we would expect the IO descriptors to be superior.

Considering the dominance between the different descriptors, it appears that one could choose one of the SO descriptors at random and produce largely the same results over a series of query images. However, choosing one of the IO descriptors would result in a system that returns slightly more varied results. This may be important if search system designers want to encourage serendipity in their search, something which is often valued by researchers.

In future work, we will pursue the evaluation of these different descriptors with users in the archaeological archives. It would be particularly interesting to know which, if any, descriptor is preferred by users in supporting their search.

Overall, the new fused descriptors proposed in this paper have a set of advantages and disadvantages over the descriptors in the SO group. In general, the new descriptors appear to work well in finding relevant and related images in an image search. However, they have the advantage of producing more varied sets of results which may add value to search systems, in particular in heterogeneous image archives where there will be large variation in material qualities of artefacts.

Acknowledgements This work was funded by the Digging Into Data Initiative through the Social Sciences and Humanities Research Council of Canada, the Arts & Humanities Research Council (UK), the Economic and Social Research Council (UK), and JISC (UK).

References

1. Archaeology Data Service, York, UK. http://archaeologydataservice.ac.uk. Accessed 9 April 2015
2. Abadi, M., Khoudeir, M., Marchand, S.: Gabor filter-based texture features to archaeological ceramic materials characterization. In: Image and Signal Processing, pp. 333–342. Springer, Berlin (2012)
3. Andrefsky Jr., W.: Lithics: Macroscopic Approaches to Analysis. Cambridge University Press, Cambridge (2005)
4. Arróspide, J., Salgado, L.: Log-Gabor filters for image-based vehicle verification. IEEE Trans. Image Process. **22**(6), 2286–2295 (2013)
5. Barceló, J.A., Pijoan, J., Vicente, O.: Image quantification as archaeological description. Bar Int. Ser. **931**, 69–78 (2001)
6. Barker, P.: Techniques of Archaeological Excavation. Taylor & Francis, Boca Raton (2003). https://books.google.co.uk/books?id=CfnHttupr28C
7. Beyer, H., Holtzblatt, K.: Contextual Design: A Customer-Centered Approach to Systems Designs (Morgan Kaufmann Series in Interactive Technologies). Morgan Kaufmann, Burlington (1997)
8. Böhler, W., Marbs, A.: 3d scanning and photogrammetry for heritage recording: a comparison. In: Brandt, S.A. (ed.) Proceedings of the 12th International Conference on Geoinformatics, pp. 291–298. University of Gävle, Gävle (2004)
9. Chalechale, A., Mertins, A., Naghdy, G.: Edge image description using angular radial partitioning. IEE Proc. Vis. Image Signal Process. **151**(2), 93–101 (2004)
10. CLAROS Project Website. http://www.clarosnet.org/XDB/ASP/clarosHome/. Accessed 9 April 2015
11. Costa, A.F., Humpire-Mamani, G., Traina, A.J.M.: An efficient algorithm for fractal analysis of textures. In: 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 39–46. IEEE (2012)
12. Digging into Archaeological Data and Image Search Metadata (DADAISM) Project Website. http://dadaism-did.org. Accessed 9 April 2015
13. Durham, P., Lewis, P., Shennan, S. (1995) Artefact matching and retrieval using the generalised hough transform. In: Wilcock, J., Lockyear, K. (eds.) Computer Applications and Quantitative Methods in Archaeology 1993. BAR International Series 598, pp. 25–30. Tempus Reparatum, Oxford (1995) **(ISBN 0860547744)**
14. Evans, A.A., Donahue, R.E.: Laser scanning confocal microscopy: a potential technique for the study of lithic microwear. J. Archaeol. Sci. **35**(8), 2223–2230 (2008)
15. Grosman, L., Smikt, O., Smilansky, U.: On the application of 3-d scanning technology for the documentation and typology of lithic artifacts. J. Archaeol. Sci. **35**(12), 3101–3110 (2008)
16. Guo, Z., Zhang, Z., Li, X., Li, Q., You, J.: Texture classification by texton: statistical versus binary. PLoS ONE **9**(2), e88,073 (2014)
17. Hesse, G.S., Georgeson, M.A.: Edges and bars: Where do people see features in 1-D images? Vis. Res. **45**(4), 507–525 (2005)
18. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **38**(12), 2270–2285 (2005)
19. Kovesi, P.: Phase congruency: a low-level image invariant. Psychol. Res. **64**(2), 136–148 (2000)
20. Kovesi, P.D.: `phasecong.m` from: MATLAB and Octave functions for computer vision and image processing. Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia. http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/#phasecong
21. Lewis, P.H., Goodson, K.: Images, databases and edge detection for archaeological object drawings. In: Sebastian, R., Lockyear, K. (eds.) Computer Applications and Quantitative Methods in Archaeology 1990. BAR International Series 565, pp. 149–153. Tempus Reparatum, Oxford (1991) (**ISBN 0860547132**)
22. Lewis, P.H., Martinez, K., Abas, F.S., Fauzi, M.F.A., Chan, S.C., Addis, M.J., Boniface, M.J., Grimwood, P., Stevenson, A., Lahanier, C., et al.: An integrated content and metadata based retrieval system for art. IEEE Trans. Image Process. **13**(3), 302–313 (2004)
23. Lin, S.C., Douglass, M.J., Holdaway, S.J., Floyd, B.: The application of 3d laser scanning technology to the assessment of ordinal and mechanical cortex quantification in lithic analysis. J. Archaeol. Sci. **37**(4), 694–702 (2010)
24. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. **18**(8), 837–842 (1996)
25. Marchand, S.: Ibisa: making image-based identification of ancient coins robust to lighting conditions. In: EUROGRAPHICS Workshop on Graphics and Cultural Heritage (GCH), pp. 13–16 (2014)
26. Marchand, S., Desbarats, P., Vialard, A., Bechtel, F., Amara, A.B., Cicuttini, B., Bost, J.P., Alain, B., Koray, K., BeurivÉ, A., et al.: Ibisa: image-based identification/search for archaeology. In: Proceedings of the 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST09), pp. 57–60 (2009)
27. Marshall, G., Dupplaw, D., Roe, D., Gamble, C.: Acheulian biface database. Archaeological Data Service, York, UK (2002). http://archaeologydataservice.ac.uk/archives/view/bifaces/. Accessed 15 Sept. 2014
28. Odell, G.: Lithic Analysis. Springer, Berlin (2004)
29. Petrovska-Delacrétaz, D., Chollet, G., Dorizzi, B.: Guide to Biometric Reference Systems and Performance Evaluation. Springer, Berlin (2009)
30. Sarfraz, M.S., Hellwich, O.: Head pose estimation in face recognition across pose scenarios. VISAPP **1**(8), 235–242 (2008)
31. Sayce, A.H.: The Archaeology of Cuneiform Inscriptions, 2nd edn. Society for Promoting Christian Knowledge, London (1908)
32. Smith, P., Bespalov, D., Shokoufandeh, A., Jeppson, P.: Classification of archaeological ceramic fragments using texture and color descriptors. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 49–54. IEEE (2010)
33. Topi, M., Timo, O., Matti, P., Maricor, S.: Robust texture classification by subsets of local binary patterns. In: Proceedings of 15th International Conference on Pattern Recognition, 2000, vol. 3, pp. 935–938. IEEE (2000)
34. van der Maaten, L., Boon, P., Lange, G., Paijmans, H., Postma, E.: Computer vision and machine learning for archaeology. In: Proceedings of Computer Applications and Quantitative Methods in Archaeology, pp. 112–130 (2006)
35. Vargha, A., Delaney, H.D.: A critique and improvement of the cl common language effect size statistics of mcgraw and wong. J. Educ. Behav. Stat. **25**(2), 101–132 (2000)
36. Vemuri, N.S., Torres, R.d.S., Shen, R., Gonçalves, M.A., Fan, W., Fox, E.A.: A content-based image retrieval service for archaeology collections. In: Gonzalo, J., Thanos, C., Verdejo M.F., Carrasco R.C. (eds.) Research and Advanced Technology for Digital Libraries, Proceeding of the 10th European Conference, ECDL 2006, Alicante, Spain, 17–22 September 2006. Lecture Notes in Computer Science, vol. 4172, pp 438–440. Springer, Berlin (2006)
37. Xie, S., Shan, S., Chen, X., Chen, J.: Fusing local patterns of gabor magnitude and phase for face recognition. IEEE Trans. Image Process. **19**(5), 1349–1361 (2010)
38. Zhang, D., Lu, G., et al.: A comparative study on shape retrieval using Fourier descriptors with different shape signatures. In: Proceedings of International Conference on Intelligent Multimedia and Distance Education (ICIMADE01), pp. 1–9 (2001)

39. Zhang, L., Zhang, L., Zhang, D., Guo, Z.: Phase congruency induced local features for finger-knuckle-print recognition. Pattern Recogn. **45**(7), 2522–2531 (2012)
40. Zhu, C., Bichot, C.E., Chen, L.: Image region description using orthogonal combination of local binary patterns enhanced with color information. Pattern Recogn. **46**(7), 1949–1963 (2013)

Mark Eramian received a combined honours degree in computer science and geophysics and a Ph.D. degree in computer science from the University of Western Ontario, London, ON, Canada, in 1997 and 2002, respectively. He has been an associate professor with the Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada, since 2002. His research interests are in the area of image processing and computer vision with a particular interest in image segmentation and an obsession with the number forty-two.

Ekta Walia is a Radiology Data Scientist with Philips Healthcare, Canada. She received her Bachelors degree in Computer Science from Kurukshetra University, India, and Masters in Computer Applications and Ph.D. from Punjabi University, India. After starting her professional career as a software consultant with DCM DataSystems, India, in 1998, she served as a faculty member in various academic and research institutions in India. She was Associate Professor and Chairperson of the Department of Computer Science at South Asian University, India, till May 2014. Later, she joined University of Saskatchewan, Canada, as Professional Research Associate to work on a research project related to archaeological image search and markup. She also has experience of conducting research pertaining to computer-aided diagnosis of thyroid cancer. Her research interests include 3D Rendering, Digital Image Watermarking, Content-Based Image Retrieval, Face Recognition, and Computer-Aided Diagnosis. She has a number of international journal and conference publications in these areas and is on the reviewing board of reputed image processing journals and conferences.

Christopher Power is a Lecturer in Human–Computer Interaction at the University of York. His research is focussed on reducing uncertainty in interactive systems, with a particular interest in digital cultural heritage archives. He works closely with users to understand their online experiences and with archive teams to improve their designs. He was the technical lead for the University of York in the DADAISM project.

Paul Cairns is a reader in Human–Computer Interaction at the University of York. He did his DPhil work in an area of mathematics considering objects that could not exist in this universe. He has an extensive interest in research methods for HCI and also the player experience of digital games. He is very fond of statistics and his favourite statistical test is Page's test.

Andrew Lewis is a Research Associate at the University of York and has a PhD. in the reconstruction of virtual cuneiform fragments in an online environment. Andrew's research interests include technology in heritage and archaeology, HCI, 3D scanning, and 3D printing.