International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 5, May-2014, pp: (377-398), Impact Factor: 1.252, Available online at: www.erpublications.com

Scalable Frameworks for Content-Based Image Retrieval

Dhiya Al-Jumeily[1], Abir Jaafar Hussain[2], Evan Brown[3], Sam Crate[4]
[1,2,3,4]School of Computing and Mathematical Sciences, Liverpool John Moores University
Byrom Street, Liverpool, L3 3AF

Abstract: This paper proposes a platform that implements search and retrieval for images based upon a query which is an example image. The paper analyses the problems of image feature identification and extraction; the performance and practicality of different types of image feature; how to match image features effectively; how to avoid false positive feature matches and thereby to retrieve accurate search results; and how to scale Content Based Image Retrieval so that it can operate on collections which contain thousands of images. The performance of different image feature detectors, descriptors and similarity computation measures is analysed. An explanation of how the most common image features and descriptors work is offered together with a prototype system for the extraction of some of these features from images. The process of matching features between images based upon their similarity will then be dealt with, firstly on a small scale in order to illuminate the matching process and then as part of a system which can grow more effectively as the number of images increases. The matching system will be systematically refined during the course ofthis focus on scale so that an efficient prototype system is ultimately presented. The issue of creating corpora of images which are suitable for search and retrieval purposes will be discussed next. A number of existing collections, their characteristics and composition will first be presented.

1. INTRODUCTION

Most mainstream image search and retrieval systems still rely on parallel streams of metadata which accompany and describe the image collection (Li & Wang, 2008). Some of this information may reference global features of the visual items themselves - information that can often be captured and stored when a photograph is taken. Some will provide a purely second-hand description of those items information which tends to provide the rich semantic detail that really facilitates finding work again. Results are retrieved from the corpus by matching terms in some way between the metadata and the user query. This only works because the two elements of content have been aligned. The problem with this approach is that creating metadata for very large image collections is time consuming and systems that support metadata production and alignment do not scale very well.

The production of metadata also generally relies upon 'late human intervention' in the publishing process. Semantically useful textual descriptions are added to a collection after the visual works themselves are created and outside the context in which that creation took place. This degrades the relevance, granularity and coverage of interrogatory activities that can be performed on the completed resource.

A significant amount of research has been published in the field of Content Based Information Retrieval(CBIR). Heightened interest here in recent years has been motivated primarily by the need for practical tools that can search the ever increasing amount of visual material on the Web. This is combined with an attempt to take advantage of more advanced mobile phone hardware and software capabilities through the delivery of augmented reality applications for game play and for the seamless delivery of location-dependent information (Wagner, Reitmayr, Mulloni, Drummond, & Schmalstieg, 2010). For example, it is reported that over one million images per day are uploaded to the Flickr publishing platform alone and search engines like Google now index billions of images (Li & Wang, 2008).

In this context, systems that are able to retrieve visual works based not upon expensively- constructed metadata streams but upon the results of an automated analysis of the information contained in the image itself are of critical importance. Good results have been obtained from testing various approaches for the computation of image similarity metrics. These theoretically enable the direct query-based retrieval of images where the search requirement is expressed not in words but in a canonical image. It seems, however, that the performance of these systems in real-world scenarios has been somewhat less impressive. Detection and accuracy rates for the underlying algorithms are good when they are run in a controlled environment against high-quality source materials from pre-prepared visual corpora. They have much more difficulty in ad-hoc situations where factors like the camera used to take query images,

the composition and aesthetic quality of those images and the nature of the environment in which the image was taken cannot be controlled.

Despite the practical and theoretical advances that have been made in the field of Content Based Image Retrieval, there is a clear requirement for further work which seeks specifically to address and evaluate the limitations of the technology as it stands. State-of-the-art knowledge has reached a position presently where end-user applications are just possible; yet the accuracy and robustness of the theory and various implementations here is not yet sufficient to support really usable tools. In fact there are distinct areas of research focus in this field which must be addressed before the technology can improve markedly and before it can become a consumer level proposition. This research work seeks to establish the state of the art in Content Based Image Retrieval, to demonstrate how the technology can be scaled to operate on catalogues containing thousands of images and more and to augment basic matching and similarity computation techniques with geometric plausibility constraints so that accuracy can be further improved and the rate of false positive matches reduced.

The final application which is envisaged in this paper is a landmark identification platform for the delivery of information about buildings and installations around the Open University campus in Milton Keynes, UK on camera-enabled mobile devices. The rationale is that this type of application of Content Based Image Retrieval technology is plausible and it could be of practical use in real-world applications. It also presents some of the accuracy problems that feature extraction and matching approaches struggle to cope with. Finally, a system for the identification of landmarks would be likely to operate on many thousands of different images and it therefore provides a test bed for the analysis of search and retrieval performance across large image collections.

2. ANALYSIS AND METHODS

2.1. Content Based Image Retrieval

The term Content Based Image Retrieval has been used to describe the process of retrieving desired images from a large collection on the basis of features, which in the early years of development centred upon general characteristics like colour, texture and shape. The retrieval of images based on keywords or textual queries does not fit within the scope of the domain, even where the words used actually describe the visual content of an image. The biggest difference between Content Based Image Retrieval and traditional textual search and retrieval is that an image database is unstructured. The only information available to the retrieval process is an array of pixel intensity values in the image files themselves. There is no inherent meaning to the content of the indexed documents. With text, the structure and meaning of any item has already been determined by the author whereas there is an overriding need with image searches to extract meaningful information from the raw image data before any retrieval activities can be undertaken (Eakins & Graham, 1999).

Many of the techniques which are applied to 12 images for content-based search and retrieval come from the broader fields of image analysis and computer vision. Some of the fundamental approaches and algorithms that are used here are also applied to problems such as robot navigation, automatic face detection, CCTV analysis and so on. There are also grey areas in the literature which overlap the boundaries between image search and retrieval and other activities. A good example of this type of hybrid undertaking is the identification of objects within pictures and photographs, as opposed to the 'pure' information retrieval task of identifying a duplicate image based upon a given visual query. In fact, some authors prefer to use the term 'near-duplicate image retrieval' rather than Content Based Image Retrieval in order to narrow the field of application and investigation in their activities.

Indeed, some authors recognise that the results obtained in Content Based Image Retrieval activities are compromised by a poor definition of what the field actually entails and what the objective of feature extraction and matching is in any given scenario. (Pavlidis, 2008) claims that there are in fact two types of search and retrieval activity that can be investigated and undertaken. The first is the general problem where we try to match a query image to an arbitrary collection of images, such as those found on the web. The goal of the search here is to obtain images which contain the same object as the query. In other words, '[g]iven an image with a horse, find all images showing a horse (at least as their main subject' (Pavlidis,2008). The second scenario involves an application specific search and retrieval task where the user wants to match a query image to a collection of images of a specific type, 'for example, fingerprints, X-ray images of a specific organ, images of skin lesions etc.' (Pavlidis, 2008). This is the less general task of duplicate detection. He goes on to claim that this problem of definition is demonstrated by the fact that many papers are vague about whether they are attempting to search for similar two- dimensional images or for images which contain the same three-dimensional object or objects. There is confusion between looking for similar images and looking for similar objects. In the case of this research work, the area of application will be duplicate image detection from a large collection of images and this will be achieved by applying Content Based Image Retrieval algorithms and techniques as appropriate.

It follows, then, that this research work does not directly address the issues which may allow for Content Based Image Retrieval to be applied better in real-world tools and applications. It is suggested that there is a stage of research which needs to be completed first in order to improve the accuracy of duplicate image search and retrieval experiments and the scalability of technical platforms before end-user requirements can be properly considered. In fact, making the jump from duplicate image and near-duplicate image detection to a search platform that would be useful in a general scenario is a very difficult problem. (Pavlidis, 2008) gives one example of the type of image search query which might be most useful in the real world.

'Find all pictures of President Clinton and Monica Lewinski' based upon a canonical image of the two people. There is no scope for a question like this to be answered by textual means because tagging generally happens at the time that a picture was taken and Monica Lewinski only became well known during the impeachment proceedings against President Clinton. The need to perform such open queries requires that we have general image similarity measures that allow for detailed matching, and possibly scene similarity measures which could provide even better discriminatory abilities. The crux of the matter is that most real-world search and retrieval scenarios do not suggest that pure duplicate image detection would be very useful. The ultimate objective here must be to enable the discovery of all images which contain particular objects. This leaves us with a complicated problem to solve, substantially because it requires a solution to a quandary known as the 'semantic gap'.

2.2. The semantic gap

Any discussion of the definition of Content Based Image Retrieval must take account of a phenomenon known as the 'semantic gap'. This gap exists between semantics and the descriptiveness of visual features. In essence, low-level pixel states must somehow be made to correspond with high-level concepts of meaning, composition, content and style (Eakins et al, ibid.). It is an apparently intractable problem unless meaning and image content –two quite different sources of data – can be united in a complimentary and practical manner. Significant effort has been expended to first understand and then address the problem here. The distinction is drawn most abruptly in semiotics between the denotation, or the presented form of an image, and the connotations or dominant impressions to which it gives rise in the mind of the beholder (Enser & Sandom, 2003).

The challenge here is to somehow unite the concepts of visual and perceptual similarity. Two pictures may differ markedly in their pixel intensity values, which are the data source that content based search operates on, but they may still appear to be similar to a human observer. On the other hand, there may be relatively small differences between the pixel intensity information for two pictures in a situation where they have very different perceptual meanings. This problem is even more severe when we consider that images may be conceptually close to one another even though they differ markedly in terms of computational similarity. Despite the significance of the semantic gap problem, however, effort is still expended to improve the perceived quality of visually-queried image retrieval results based on the comparison of pixel data. This is thanks largely to a focus on the manner in which and selectiveness with which image features are chosen, tested and stored as somehow significant content signatures which differentiate or categorise a visual work.

The phenomenon of the semantic gap has been described as being closer to a 'semantic abyss' because of the apparent lack of suitability of image data as a descriptor of perceptual and particularly conceptual similarity (Chiu, Lin, & Yang,2003). In text files, words are encoded as strings of bytes which represent character codes. These groups of character codes have direct meaning which is obvious to a user when the text is reproduced. Text can be quite easily processed, evaluated for similarity and ranked using linguistic principles and natural language processing techniques. These techniques may include collocation, term frequency-inverse document frequency (TF-IDF), Mutual Information and so on. It is worth saying here, however, that even with text, complex queries about the semantics of information are still hard to do. A great deal of effort is being expended to enable information retrieval activities which are concerned with the meaning – and the inter-related meaning or significance – of text data. The development of ontologies and frameworks for mapping conceptual meanings on to words seeks to facilitate more detailed parsing of written information and online developments like the semantic web (Thiagarajan, Manjunath, & Stumptner, 2008). At present, however, it is still problematic to answer the query: 'Find all critical articles about the coalition government which relate to its policies on the welfare state.'

In the domain of image retrieval, there is ultimately a need to replicate – or at least to imitate in a convincing manner - complex processing and contextual transformations of information which have evolved in the human and animal visual systems over hundreds of millions of years. It is also worth saying that visual processing of image data is not one dimensional or unidirectional. Contextual information that is gathered from both inside and outside the image by the human brain helps us to recognise what we see. Without this contextual information, it would be impossible to tell the difference, for example, between a scale model of a London bus in one photograph and a real London bus in another, despite the fact that the presentation of either object as a visual query could require the retrieval of very different information in response. It must be recognised that the general Content Based Image Retrieval problem cannot be

solved without significant advances in the fields of image analysis and computer vision. As Pavlidis proposes that effort should be expended on solving specific image retrieval problems which satisfy certain criteria: that the mapping between semantics and image features is well defined; that top- level knowledge and context for the search scenario is known; that accuracy requirements are well defined; and that the matching of images requires careful scrutiny of the pictures involved (Pavlidis, 2008).

2.3. Image features

Extracting features from images which can be said to be diagnostic of the work as a whole is a challenging process, partly because of the amount of information that is encoded in image files. Even a simple grayscale picture can contain millions of pixels, each of which is represented by an intensity value of between 0 (black) and 255 (white). Therefore one of the prime initial tasks in Content Based Image Retrieval is to make some sort of sense out of this plethora of numbers. The key here is to condense the amount of information in an image down into meaningful and robust pieces of information. These pieces of information are called features. The process of indexing images can be seen as one involving the creation of summary statistics which describe the image at a low level. These statistics may represent colour usage, texture composition and shape information and spatial structure, amongst other phenomena. They may also represent some more complex 'local' features of an image which are based on gradients and intensity changes in different small regions of a picture. In general, a large number of feature extraction techniques have been proposed and investigated but there is still no satisfying general solution to the problem of information extraction from images (Deselaers, Keysers, & Ney, 2008).

The simplest approach to Content Based Image Retrieval is to directly use the pixel values encoded in an image file as features. Here the pictures must be scaled to the same size and the Euclidean distance between individual pixel values can then be used to calculate similarity. However, even a simple image can contain many millions of pixels which each have multiple associated pieces of information to denote intensity and colour. This is therefore an inefficient and computationally intensive method for comparing images. One way to improve efficiency would be to scale down the query and catalogue images to a relatively small size, although lot of information and significant discriminatory characteristics are naturally lost as a result of this process. Even where images are scaled down, however, it is likely that the approach would not scale very well to medium or large image collections (Deselaers, Keysers, & Ney, 2008).

Another more efficient approach is to use colour histograms as image features. These provide a crude but ultimately diagnostic summary of the colour usage in an image as a whole. The colour space is partitioned into colour bins and the proportion of pixels in the image which fall into each bin is calculated. Colour histograms are often applied to the red, green and blue (RGB) colour space as this equates well with the peaks of colour sensitivity in the eye receptors of the human visual system. It is possible and quite common to use the same approach in different colour spaces however. Hue and saturation based models can also be used, where variation is used (HSV), where luminance is used (HSL) and where brightness is used (HSB). The colour histogram for an image is thus represented as a vector of real numbers which are calculated by partitioning the image into cells and counting the number of pixel occurrences for a particular colour in each cell. This vector is then normalised and distance calculations between similarly-encoded vectors (different images, in other words) can be undertaken (Little, Brown, & Rueger, 2011).

2.4. SIFT and SURF image features

Perhaps the best known 'local feature' and most-used system for feature extraction and description is the Scale Invariant Feature Transform(SIFT) (Lowe, 1999) (Lowe, 2004). This is an unusual proposal because it specifies not only a feature detector but also a format of descriptor as well. As has been noted by some commentators, SIFT key points can be used in combination with other detectors. They tend to describe blob-based regions of an image whereas other techniques based on the Harris scheme describe edges and corners. Therefore an amalgamation of the two approaches can provide for a rich description of image detail. In general, SIFT is recognised as a very robust and reliable scheme for interest point detection and description. It is based on simple theory which is relatively efficient. That being said, SIFT feature descriptors are composed of a high number of dimensions and they can be computationally intensive to calculate.

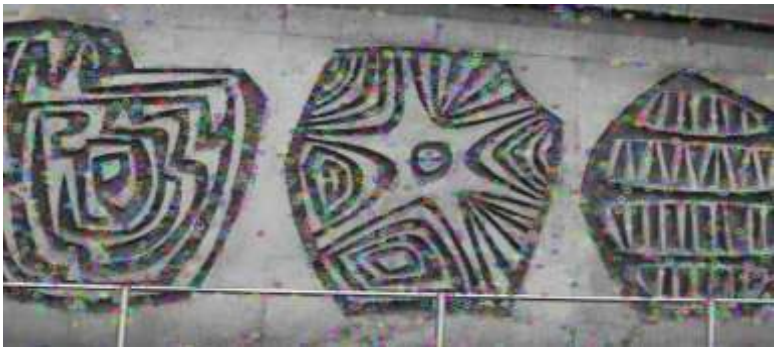


Figure 1: SIFT features detection in a photograph

Given their relative robustness and high levels of performance, SIFT features have been used extensively for image analysis, matching and object recognition tasks as shown in Figure 1. However In fact, there have been several attempts to improve upon the performance and accuracy of SIFT features and matching systems for specific applications. (Bay, Tuytelaars, & Van Gool, 2006) propose Speeded Up Robust Features (SURF), which are significantly simpler and quicker to compute than SIFT and which consume far less storage space because the local descriptor vectors are smaller in size. In fact, SURF descriptors are composed of 64 dimensions as opposed to the 128 dimensions of a standard SIFT descriptor. As well as being easier to calculate, SURF was said to offer slightly better retrieval accuracy than the older system. (Bauer, Sunderhauf, & Protzel, 2007) did not find this to be the case – although the difference in matching performance was minimal and still impressively high with the SURF algorithm – and it is noted that the principal advantage of Bay's system is the reduced computational requirements that it brings. The authors also stated that, whilst SURF tends to highlight fewer significant key points in an image than SIFT does, this in itself did not seem to adversely affect retrieval performance and can be seen as an advantage in most circumstances. Perhaps more interestingly, however, (Tola, Lepetit, & Fua, 2010) noted that SURF features generate artefacts that degrade the matching performance when used in images with dense key points. Figure 2 shows the SURF features detected in a photograph.

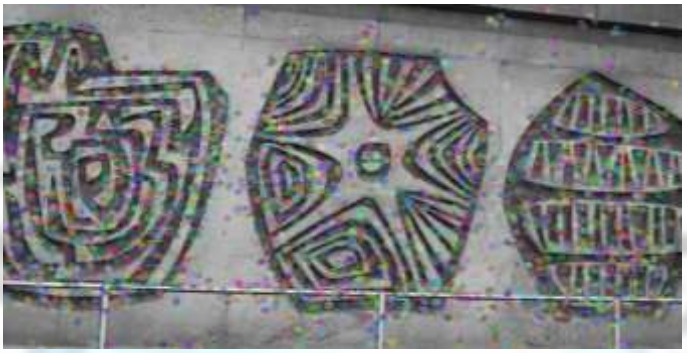


Figure 2: SURF features detected in a photograph

2.5 Proposed bag-of-words system

A prototype system for matching images will be presented which implements a bag-of-words scheme for document codification based upon the creation of an initial visual vocabulary from a corpus of images. The prototype system uses the OpenCV library for feature detection and description purposes. A decision has been taken here to use a SURF descriptor and detector framework rather than the SIFT scheme that was employed previously. SURF provides reduced dimensionality for feature vectors that are detected in images, with sixty four values per vector rather than one hundred and twenty eight, this helps to ameliorate the storage and memory requirements which are implied when dealing with a large collection of images for indexing and searching purposes.

The SURF detector is also faster at initially identifying regions of interest within pictures and the fact that it does not detect as many key points as the SIFT approach is considered later when addressing the retrieval accuracy of the platform across different collections of images. This provides for the creation of a visual vocabulary from a collection of image features which have been extracted from a corpus of images. The visual vocabulary is defined through clustering the features, as discussed previously, and then using the centroids of these clusters as visual words which can represent classes of features that appear in pictures from the collection and, ultimately, in novel query images which are dealt with by the system.

The prototype system consists of three separate components: an indexer module which processes a collection of pictures, extracts feature vectors from them and then creates the initial visual vocabulary for search and retrieval; a server component which provides access to the image collection and performs search and retrieval operations on the collection in response to properly formatted query requests; and a simple client application which provides a web-based user interface to the system and allows for the uploading of query images and the creation of queries against the collection before presenting search and retrieval results to the user.

3. RESULTS AND DISCUSSION

3.1 The Open University Art collection (OU Art)

The novel bag-of-words platform which was shown earlier will be used for image retrieval. The first set of tests will be performed using the small Open University Art (OU Art) picture collection. This collection contains eighty one images of various subjects. Some of the pictures feature multiple viewpoints of the same object although this was not a design imperative for the corpus and such repetition is not reliably repeated throughout the collection. The collection will be

interrogated manually using a set of query pictures that have been taken on a camera enabled mobile telephone. This query set has been specifically produced for the purpose of testing retrieval accuracy and the pictures feature different shots of the same objects which appear in the collection itself. There are a total of thirty two query pictures which consist of different viewpoints of some of the objects and items of interest that have been captured in the corpus. The nature of the query pictures raises an interesting question about the effect of image resolution on the accuracy results which can be obtained using local image features as the basis for matching and similarity assessment between different pictures.

In the initial instance, a basic score for the percentage of pictures in the query set which correctly match an image in the corpus can be obtained. The matching is done manually by presenting each available query to the search and retrieval system through the simple web front end and recording either the success or failure of the interrogation request. The system is judged to have matched the query correctly to the collection if an appropriate image response occurs anywhere within the top four results that are presented to the user. Using this approach, it has been found that the system correctly matches 93.75% of the query pictures to items in the collection and this represents a retrieval performance where thirty out of the thirty two available queries pull back correct pictorial results.

There are a couple of things that are interesting to note here. Firstly, the correct item in the collection was almost always the first ranked search result in the set of retrieved images where a correct match was indeed found. Only one query produced a situation where the appropriate result was ranked lower than first and this was an image of a plaque. One picture of a plaque had indeed been retrieved as the first result but it was a different item to the correct choice which was ranked as search result number two. It is suggested that plaques and similarly inscribed objects make for a difficult retrieval proposition because the writing on them is usually too small or indistinct to generate good features that can be tracked between pictures. The situation could be different if the photographs of these items were taken close up so that the scale of the text in the picture is large and the proportion of the image taken up by the inscription is significant. However, the plaque pictures in the OU Art collection are taken from some distance away and, although it is possible for a human viewer to differentiate the different pictures in this class, it is likely that a text recognition algorithm or a pattern matching approach would be required in order to significantly boost performance here.

The second thing which is worth mentioning is that matching performance was universally poor where a correct match did not appear as the first returned image. In other words, none of the pictures that were deemed to match a query were actually correct if the first item in the list was an incorrect match. This means that the ranking algorithm is being fed poor quality information in these cases and it somewhat suggests that the mapping of image features to centroids in the visual vocabulary completely breaks down in the majority of cases where matches were not clear cut. One might expect to see search result listings where the first item was not correct but the second, third and fourth results contained at least an image which was some viewpoint of the object of interest in the query. This is not borne out, however – at least, not with the small OU Art dataset.

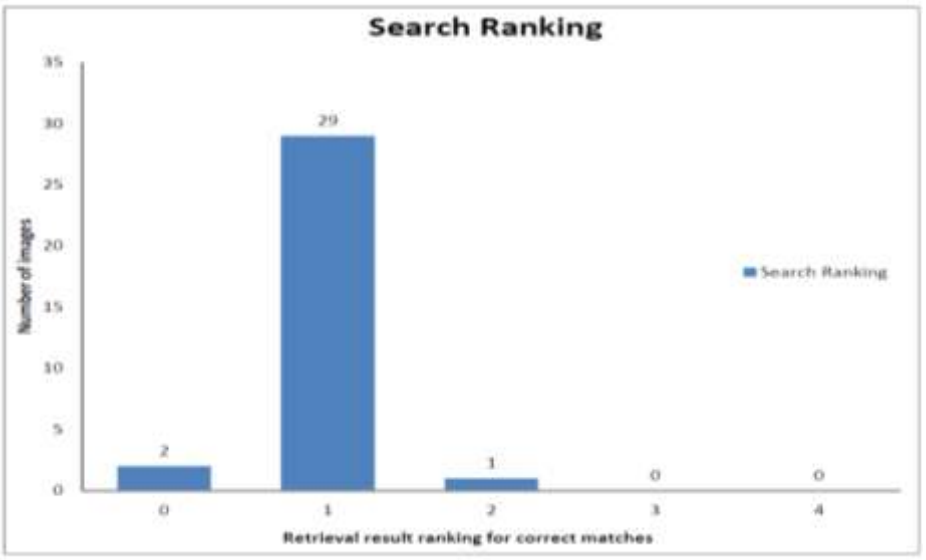


Figure 3: Retrieval rank for correct matches generated from the OU Art

Figure 3 demonstrates the search rankings that were obtained using the search and retrieval system with manual querying and result analysis on the OU Art dataset. A search ranking of zero in this context means that an incorrect

image was returned in response to a given interrogation request and that no correct image was shown in the top four search results for that particular query. Consider the precision of the search which is defined as:

$$\text{Precision} = \frac{\text{Number of relevant records received in response to a query reques t}}{\text{Number of relevant records plus the number of irrelevant records that were returned}} \quad (1)$$

The precision graph for the Open University art collection is as shown in Figure 4.

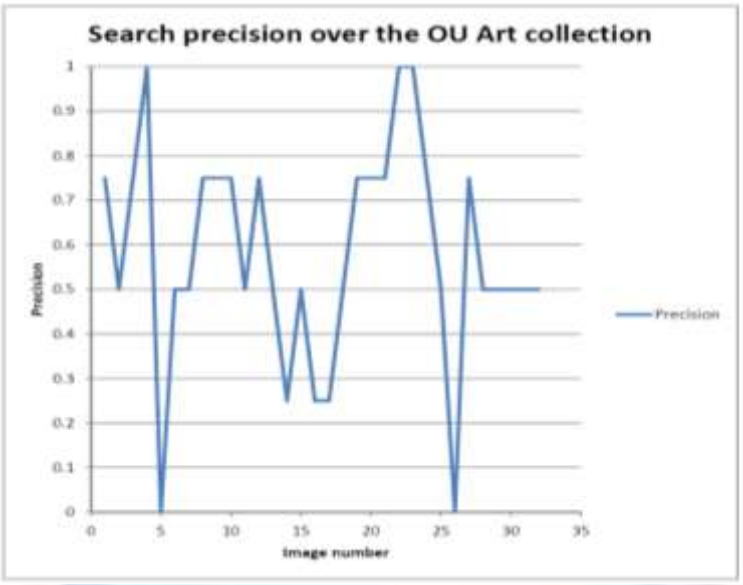


Figure 4: Precision results for retrieval of images in response to the query

The average precision rating for search and retrieval operations across the Open University art collection is 0.578, or 57.8%. In a textual search environment, the searcher can use different strategies for expressing their query in order to boost the precision level of the record set that is returned. In the image retrieval scenario, this is harder to achieve since the parameters of the query are controlled largely by the image features that are extracted from a given picture and the method in which these are matched to centroids in the vocabulary in order to express the content of a particular picture. There is a distinct semantic gap here between the understanding that a user will have about the material that they are looking for and how this requirement can be translated into meaningful regions of interest which are based on abstract principles such as gradient direction and intensity changes at the pixel level. Textual search results can be altered simply by adding tokens to the input query or by expressing the information requirement in a different way but it is much harder to alter the information request in an image search context. The feature detection and description algorithms and the index creation process all accept a number of different parameters which control the manner in which they are constructed and the effect of changing these values may equate to imposing a different search strategy on the visual material that is available to the user. Unfortunately many of the parameters which control how a corpus is initially analysed and described become hard coded because the index preparation step cannot be completed prior to each novel query. It becomes fixed once a collection is processed and the database files are created. Re-indexing a collection is a major step which is resource intensive especially for larger collections of pictures. It is not feasible to repeat this process prior to each different query and, in any case, it is not clear what the relationship is between the different parameters that can be set here and the way in which they express a particular information need on behalf of the user. The query picture can be analysed differently prior to each interrogation because features are extracted from this resource in real time before the index is queried to pull back matching images. However, the semantic gap problem still applies here and the fact that the visual vocabulary into which each image is transposed already exists meaning that the scope for really altering which results will be retrieved in response to a given query is limited.

In any case, the result of this analysis shows that the search and retrieval system is accurate on the OU Art collection but that it is not particularly precise. This means that the platform often matches a query image to the correct picture in the collection but that the search results which appear lower in the ranking than the first placed item are not relevant or related to the query image. It would be desirable to achieve a situation where the search results were ranked elegantly, meaning that appropriate images of the same item of interest are returned in some sort of order in response to a particular query image. Part of the problem here may be that the OU Art collection is small and does not contain a high level of repetition in the items which are captured.
A different collection such as the UK Bench corpus might display a higher level of precision because objects are grouped into sets of four pictures and all four of the images within each group will be relevant to a given query. This is

partly because the UK Bench scenario employs query images which are taken directly from the collection and, as such, they do not exhibit problems such as blur; occlusion, perspective and viewpoint change from the other items which they should match. In that sense, results from the OU Art corpus are interesting because the situation where a user takes a picture of an item of interest on their own cameraenabled device and then uses that to ask for matches is closer to the real world application for which content-based image retrieval tools are likely to be used.

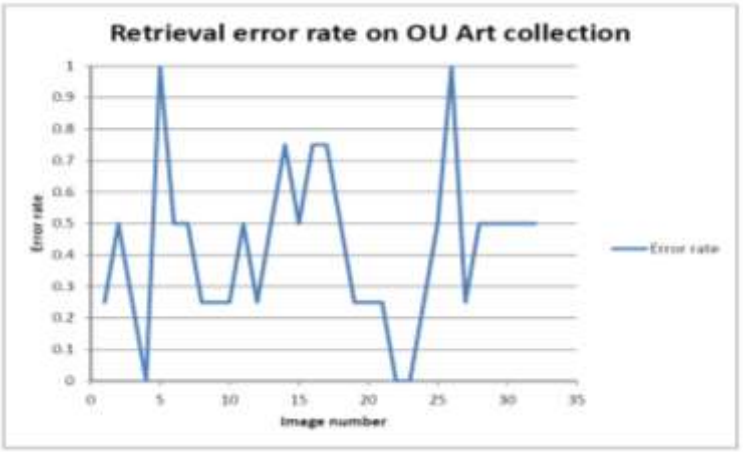


Figure 5: Retrieval error rate on the OU Art image collection

From the preceding discussion and the error rate graph which is presented in Figure 5, it is clear that the search and retrieval platform displays a relatively high number of matches in response to each query which are incorrect. This does not affect the overall accuracy of the system in this case because the correct image is displayed at search rank one in most cases. However, the subsequent series of results that is shown after a correct match often contains irrelevant images even where there is a degree of repetition in the corpus which should ensure that similar images to the primary match are available and should be able to be shown to the user in theory. The statistics for the error rate which is evidenced in response to each query image equate to an average error rate across the collection of 0.42 or 42%. This means that 42% of images that are shown in response to interrogation requests are inaccurate. As previously stated, it may be possible to reduce this metric when the system is used to index and search the UK Bench corpus because this collection breaks down neatly into groups of four pictures which are all relevant to specific image queries from the same set of pictures.

3.2 UK Bench Dataset

It is now desirable to measure the search and retrieval performance of the bag-of-words system on the UK Bench dataset. This corpus consists of a total of 10,200 images which can be broken down into smaller sets in order to measure the relative performance of the system on increasing numbers of images. The percentage accuracy which is achieved by the platform will be measured over image sets up to the full size of the complete corpus. The average number of images that is returned per group of four relevant pictures will be calculated and shown. Precision ratings will be extrapolated from the statistics using the average precision which is obtained over the same sets of one thousand and twenty pictures. The number of relevant images returned over the complete collection will be demonstrated as a histogram and the error rates achieved on average will be presented.

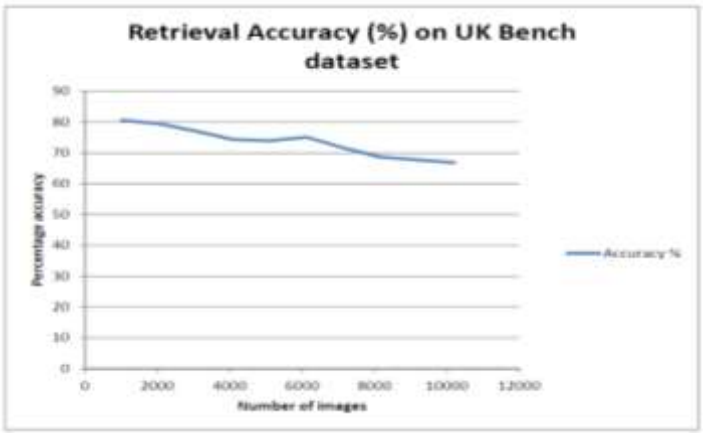


Figure 6: Accuracy percentage on the different sets of the UK Bench collection

As can be seen from Figure 6, retrieval accuracy using the bag-of-words system generally decays over the UK Bench collection as the number of images which has to be searched increases. This phenomenon is not really a function of the increasing number of images in itself. The decline in accuracy is to be expected due to the design of the corpus. Later sets of images tend to be less well defined than those in the early part of the collection. Geometric shapes and items of interest with clear areas that will give rise to regions for local feature identification and description give way to objects that are harder to differentiate from different viewpoints and perspectives. The obtained results have been constructed from increasing numbers of image sets from the corpus.

Readings were taken for every group of one thousand and twenty images that go to make up the corpus and the collection size was cumulatively increased according to these steps until the full corpus of ten thousand two hundred images was finally interrogated. These ten degrees of decomposition equate roughly to the boundaries of different sets or types of images in the collection. The early images feature items such as jigsaw boxes and compact disc covers which the authors note are a relatively easy matching proposition. Later pictures include photographs of people in different environments and smaller objects which are more complex or which take up less space within the image, and are therefore harder to extract meaningful local features from which do not become swamped by features in other incidental elements of the shot.

The accuracy of the system on the first subset of 1,020 images is 80.6% and this reduces to 66.8% when measured across the full corpus of 10,200 pictures. The accuracy figure for five thousand one hundred images increases slightly from the general downward trend and the improved performance here is carried through to levels for subsequent subsets of the collection, albeit that the tendency to decay resumes from this point. This behaviour can again be attributed to the composition of the collection, where clear and unambiguous pictures of objects are interspersed somewhat with more complex matching propositions. It is not clear to what extent the pictures were taken with the intention of progressively increasing the difficulty of the matching problem. It is hard to construct a corpus which presents a uniform level of difficulty to feature extraction, description and matching algorithms because different items in the natural environment are necessarily easier or harder to identify even for human viewers.

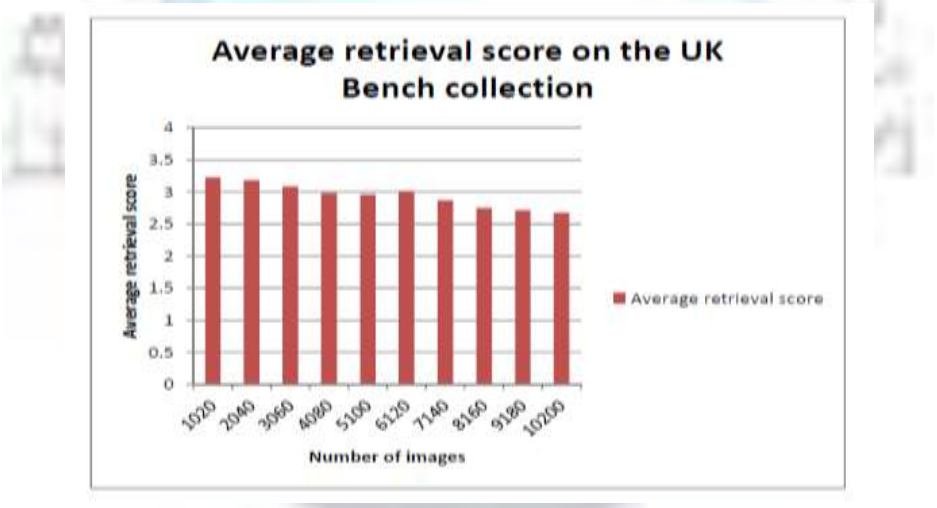


Figure 7: Average retrieval score in terms of image sets on the UK BenchCollection

Figure 7 shows the average retrieval scores which were achieved on the different subsets of the UK Bench corpus. This metric is based upon the fact that there are four pictures in every group of the same item or object and any picture in a group can match to all the others in the same group, including itself. Therefore a single picture can score up to four points where it is correctly matched to all the other pictures in the same group as itself. The average retrieval score across different subsets of the whole corpus peaked at 3.22 on the first set of one thousand images. This corresponds to the group with the highest number of simple object photographs from a matching perspective, featuring boxes and other geometric shapes which provide good material for local feature detection. Most of the images seen here feature the object of interest prominently in the frame of the photograph and so there are many features which describe and can track the item from one picture in the group to another. The lowest average retrieval score under this metric is seen on the full corpus of ten thousand two hundred pictures, where it was reduced to 2.67.

This means that just over half of the matches that should have been found with ideal performance levels were indeed identified and just less than half of the images retrieval score returned where irrelevant to the query. The number of correct matches that were recorded for query images from the subsets of the collection which have been used here can also be visualised as follows in order to give further information:

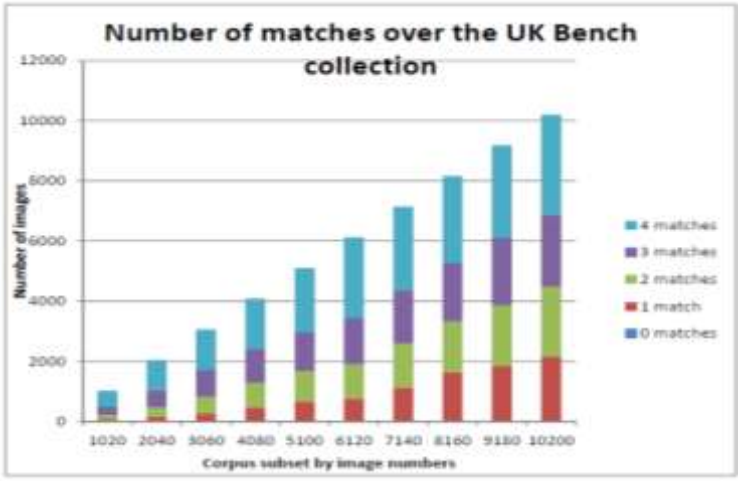


Figure 8: The number of matches per query image over subsets of the UK

The most interesting fact which comes to light from the graph shown in Figure 8 is that there are very few query images from the collection which retrieve no correct matches at all. Almost all of the pictures that are used as interrogation requests pull back at least one image from the same group as themselves. In fact, only four images lead to the display of completely irrelevant information when used as queries over the corpus as a whole. This represents excellent performance when it is considered that the search and retrieval system was limited to bringing back four results for each query, which means that for all images apart from four, a relevant picture was top-four in the search rankings. The other point that can be made about the above graph is that there is lead to the retrieval of a complete set of relevant images. This is particularly true in the first subset of 1020 pictures, where 531 queries resulted in an entirely relevant set of matched photographs. The proportion of pictures which lead to the retrieval of less than four correct pictures gradually increases as the corpus size grows and this is to be expected given the comments about object complexity which have already been made.

Nevertheless, a total of 3312 pictures from the complete collection were matched entirely correctly to the four other pictures in their groups and this represents 32.5% of the queries. By contrast, the number of pictures which did not produce any correct matches remains static over the corpus once they are introduced in the six thousand one hundred and twenty image subset. The fact that this number does not grow also indicates that the search and retrieval platform is performing well. By the end of the analysis, the fastest growing score banding is the two match's category indicating that the overall performance of the system is moving down towards 50% accuracy. In fact, as has already been shown, the overall accuracy over the full collection is some 66% and the proportion of pictures which still match all the correct images in the collection is still growing and significant at this point. Given the scale of the collection and the differing levels of difficulty which each image and each subset of images heralds, it is suggested that this represents an acceptable level of performance without the use of additional technologies such as Random Sample Consensus (RANSAC) model fitting and blur detection to help with the identification of objects of interest and the portions of each picture which should be considered as important for the extraction of local image features.

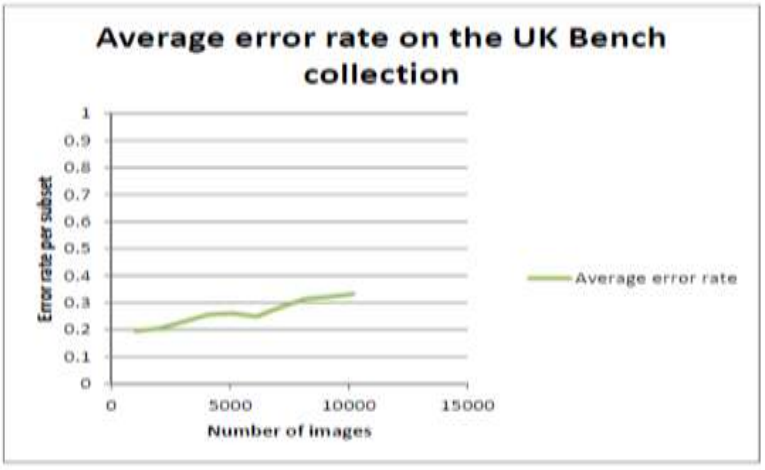


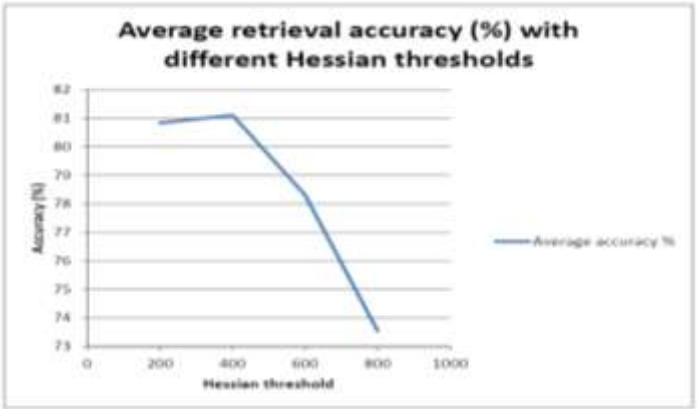
Figure 9: The average error rate across subsets of the UK Bench image Collection

Finally, it is possible to consider the average error rate which can be achieved using the search and retrieval platform on the UK Bench dataset. This metric evaluates the number of irrelevant images which are pulled back in response to each query and it averages that number across different subsets of the UK Bench collection as the overall size of the corpus grows. The graph shown in Figure 9 is inverted when compared to the accuracy and precision visualisations which have been shown previously and it reflects the fact that errors are less common in the early subsets of the collection but that they grow in frequency in the later elements of the corpus. This may once again be down to the design of the collection and the fact that the difficulty of the matching task appears to grow as more images are considered. An error rate of 0.19 is recorded on the first subset of one thousand and twenty images and these heavily feature boxes, packaging and compact disc covers. The rate of error grows steadily from this point and it totals 0.33 when the full scope of the ten thousand two hundred images is finally considered. A thirty three percent error rate may seem fairly significant but it is less important in cases where the first picture in the search rankings is a correct match and subsequent images which are deemed to be less relevant are in fact wrong. Previous analysis suggests that this scenario arises frequently throughout the collection because the number of false positives very rarely translates into a series of four images where none of the corresponding matches are in fact correct. In this respect, the UK Bench collection does not really represent a good test of actual search and retrieval utility because the situation where a user is looking for matching pictures from a set based upon a query which is part of that set would not arise very often.

Certain applications of content-based image retrieval, such as the discovery and analysis of medical images such as x-rays, may implement a similar matching platform but the basis of this report is to look at how users would be able to query a collection using their own captured visual material in order to find more pictures of the same and similar objects as have been photographed. Indeed, comparing the results from the OU Art collection to those that have been recorded for UK Bench indicates that the system performs better in terms of both accuracy and precision in the first scenario, where query pictures have indeed been captured on a camera enabled mobile telephone. This fact is unlikely to be explained purely by the smaller size of the first corpus, although it is worth reiterating that the visual vocabulary that was generated for OU Art is comparable in size to that produced under UK Bench and this may provide more discriminatory information in the former case. Nevertheless, the indications are that the system performs well when presented with truly novel query images which provide different representations of items of interest that are to be found in the collection. It may in fact be true to say that the pictures in both the query set and the collection set for OU Art are less complex or difficult from a feature detection and description standpoint than some of the pictures that are to be found in the UK Bench collection.

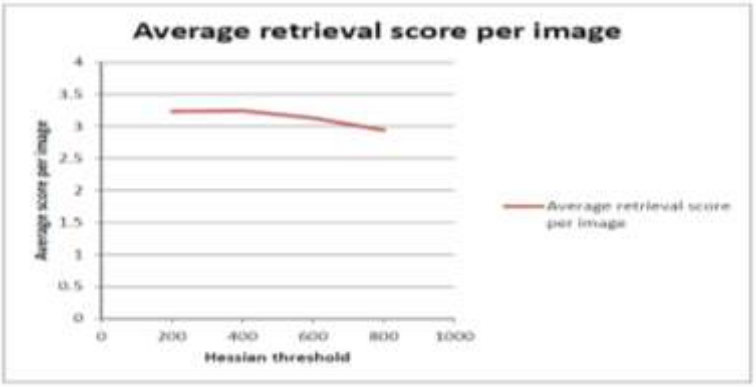
3.3 The effect of Algorithm Parameters

It is possible to conduct experiments which evaluate the effect that a reduced and then progressively increased hessian threshold has on the accuracy of the bag-of-words retrieval system by taking the normal default value for this parameter and then altering that value around the default. Reducing the hessian threshold has the effect of increasing the number of key points which are detected within an image because the intensity response of different regions can be lower than normal and the system will not discard them at the same rate as it would do with a higher threshold value. Conversely, increasing the hessian threshold raises the level of intensity response that is required in order for an interest point to become qualified for further processing and so the number of key points which are detected in the image is reduced. Whilst a general rule states that more key points identified in a picture leads to better matching performance, this is true only to a point. It is important to ensure as far as possible that the quality of the points which are detected is good since many identified key points which are localised to artefacts in the image or structures other than those associated with the object of interest do not contribute effectively to the recognition problem. Therefore the hessian threshold parameter can be used to disqualify weak points and the response strength within a region is one indicator of the quality and reliability of the feature that has been detected. A strong feature provides a good candidate which can be matched across different versions of the same image with accuracy, whereas a weak feature tends to be transitory or at worst can confuse the retrieval process by presenting false matches to regions in other pictures.



Average retrieval accuracy over a subset of the UK Bench corpus with different Hessian thresholds for the SURF algorithm

It is worth noting here that the accuracy of retrieval operations peaks when the default hessian threshold value of four hundred is used to prime the algorithm. Here, eighty one per cent of images are matched correctly to images in the UK Bench collection. It should be stated that only a subset of the full corpus has been used here but this has been selected to include those pictures which generated the highest accuracy ratings in previous measurement exercises. Accuracy also reduces with increasing values of hessian threshold and it drops away to seventy three percent when a threshold value of eight hundred is used. These findings suggest that the default value for the parameter has been chosen because it provides the best results in most scenarios. Nevertheless, it appears that increasing the threshold from four hundred in this case has not just disqualified weaker interest points which are incidental to the matching process. It is clear that many of the regions of interest which do not pass the stricter thresholding test are critical to accurate matching performance and they must denote structures which are important in the pictures that are being compared.



Average retrieval score per image over a subset of the UK Benchcorpus with different Hessian thresholds for the SURF algorithm

The above graph shows the degradation in retrieval performance that is evidenced when the hessian threshold parameter is altered and, once again, scores drop off regardless of whether the value is set lower or higher than the default that is used by the SURF algorithm. An average score of 3.2 per query image is achieved with a threshold of two hundred and these increases to 3.24 when the default value of 400 is implemented. Increasing the hessian threshold further results in a faster drop off in performance, with a score of 3.1 per image being recorded for a threshold of 600 and a score of 2.94 per image being recorded with a value of eight hundred here. Once again, the implication of this finding is that important key points within each image are progressively lost because they do not meet the requirements for an increased intensity or strength threshold which is put in place when the hessian threshold parameter is increased beyond the default. The fact that performance suffers when the threshold is lowered means that additional key points are detected within a picture which are not useful to the matching process because they do not describe features which are integral to the object of interest and therefore serve only to produce false matches between the query image and candidate pictures within the collection. The precise effect of raising and lowering the hessian threshold can be seen in the following table, which summarises the number of key points that are detected in one image based upon the level of the qualifier that is used here.

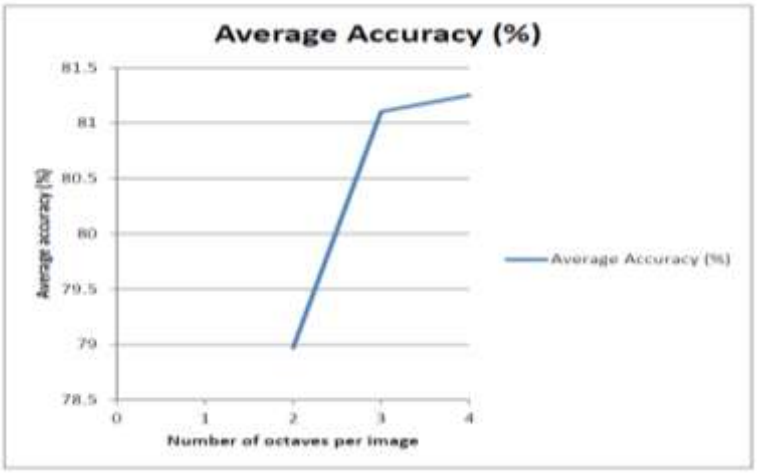
Hessian threshold	Number of keypoints detected in sample image
200	1794
400	1261
600	1011
800	866

Sample image from the UK Bench corpus used for hessian thresholdevaluation, with SURF features shown
Another parameter which is accepted by the SURF algorithm and which can be changed in order to affect the accuracy results which are obtained here is the number of octaves that the system processes for key point detection. In the case of the SURF algorithm, this relates to the number of differently sized filter groups which are applied to the image in order to record intensity response results from the underlying data when the box filters and the integral image approach is applied to the underlying pixel data. The default number of octaves which is used for processing images is three. Higher octaves use larger filters and they subsample the image data. Increasing the number of octaves here will lead to the discovery of larger regions of interest or blobs within the picture whereas reducing the number of octaves means that larger key points will not be detected but smaller regions of interest will be shown preferentially.

The initial assumption here is that a larger number of octaves will produce more stable matching results since it will lead to a bias in favour of detecting and describing significant image regions which are readily visible and which themselves lead to differentiation and identification of one image as opposed to another by human viewers. The counterpoint to this theory states that detail will be lost in the detection phase because the larger filters that are applied to an image will produce fewer blob responses at a larger scale than a smaller set of more finely applied box filters can furnish. The effect of varying the number of octaves which are processed from the same test picture that was used before is summarised in the table that follows.

Number of octaves	Number of keypoints detected in sample image
2	1184
3	1261
4	1266

The results show that increasing the number of octaves which are processed on an image tends to produce more key points. On the basis that a greater number of detected and described regions of interest leads to more content with which to match a picture to other similar items in the collection, it might be expected that accuracy results improve as the number of octaves is progressively increased. This turns out to be a well-founded assumption, as the following graph of accuracy results demonstrates. It should be noted here that a subset of the UK Bench corpus was used for this exercise and this happens to correlate with the first one thousand and twenty images in the corpus, since these produced the highest accuracy figures in previous experiments. The results were obtained by processing the subset of pictures during the indexing phase with a progressively higher number of octaves. Thus the collection was re-indexed each time an increase in the number of octaves per image was required. Each image in the group was then used as a query on the small collection and feature extraction from this picture was also parameterised and amended so that the number of octaves that were processed from the query matched that which was applied to each image in the collection as a whole.

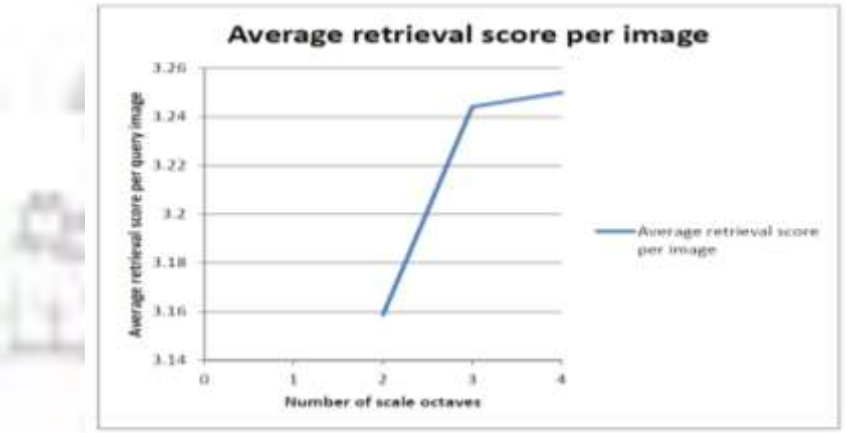


Average retrieval accuracy over a subset of the UK Bench corpus as the number of octaves extracted per image increases

The results here show a relatively sharp increase in average accuracy as the number of octaves to be extracted per image is increased from two to three. The accuracy figure starts out at 78.9% with two scale octaves and this figure increases to 81.1% when the default number of three octaves is applied. The best accuracy figure is achieved with the highest number of octaves that has been tested, and this stands at 81.25% with four scale octaves per image. This observation should be balanced against the fact that increasing the number of octaves that are considered also makes the computation time for each image longer as the feature detection algorithm has to apply more groups of box filters to each picture in turn. The default value of three octaves has probably been chosen by the designers of the SURF

algorithm to provide a reasonable balance between distinctiveness, accuracy and computational efficiency. It is also not clear from the results that increasing the number of octaves for the SURF algorithm to work on would improve retrieval accuracy in all situations. It may be the case that some images benefit from more analysis at larger scales whereas others do not exhibit the same tendency.

Large blobs or regions of interest are likely to correspond to significant elements of the picture such as edges, contours and corners which help to differentiate the item that must be identified from the background and smaller structures that are ultimately extraneous to the matching problem. It is therefore likely that increasing the number of scale levels through which the image is processed helps to reduce the relative importance and impact of smaller artefacts which give rise to features but which do not denote important regions within the image from the standpoint of identification and matching. The sample image which is shown with features added demonstrates the point here. It highlights a multitude of small regions of interest particularly in the bottom right hand corner of the picture which relate to gradient changes that can be seen in the background fabric upon which the item of interest has been placed in order for the photograph to be taken. There are in fact more features shown here than on the toy itself and these responses are only useful insofar as subsequent pictures in the same group demonstrate equivalent amounts of the background in comparison to the object. The prioritisation of larger image features probably leads to more blob responses in the body of the toy and these are clearly more useful because they are distinctive and help to delineate the features of the object itself which can then be matched more easily to subsequent versions of the same scene. The same trend for improved scoring results as the number of image octaves is increased can also be seen in the following graph which considers the average score of each query image in the first subset of the UK Bench corpus.



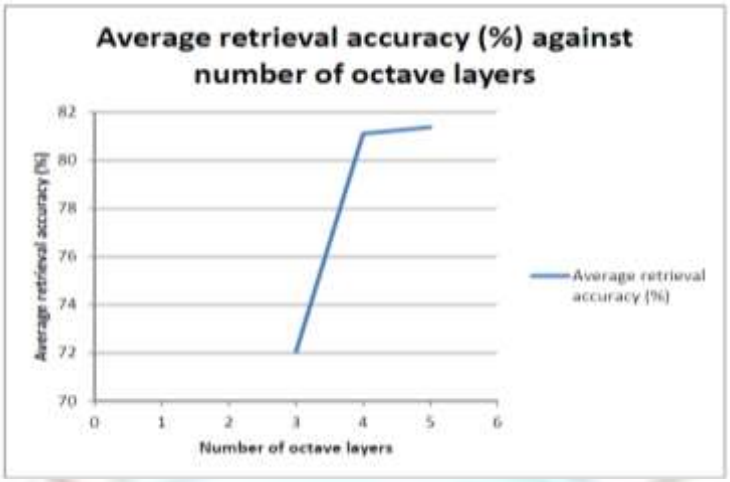
Average retrieval score per image on a subset of the UK Bench corpus with increasing numbers of scale octaves in the SURF algorithm

The next step here is to find out what the effect of increasing the number of scale levels per octave is on first key point detection and then on retrieval accuracy. With the SURF algorithm, this equates to adding additional box filters of various sequential sizes to each octave of the scale space so that the overlap between different scales happens at a different point and the intensity results or blob responses are more finely analysed over the space as a whole. An increased number of levels per octave therefore has the effect of identifying more regions of interest at finer scale increments. The default number of scale levels which are considered per octave with the SURF algorithm is four. The results of an analysis of the number of key points which are detected in a sample image with increasing numbers of octave layers to prime the algorithm are given in the following table. The sample image that was used for analysis here is the same one that is shown previously. In order to derive these results, the number of octaves used by the algorithm has been set back to three and this level of analysis is maintained whilst the number of levels per octave is increased progressively.

Number of levels per octave	Number of keypoints detected in the sample image
3	1158
4	1261
5	1356
6	1443

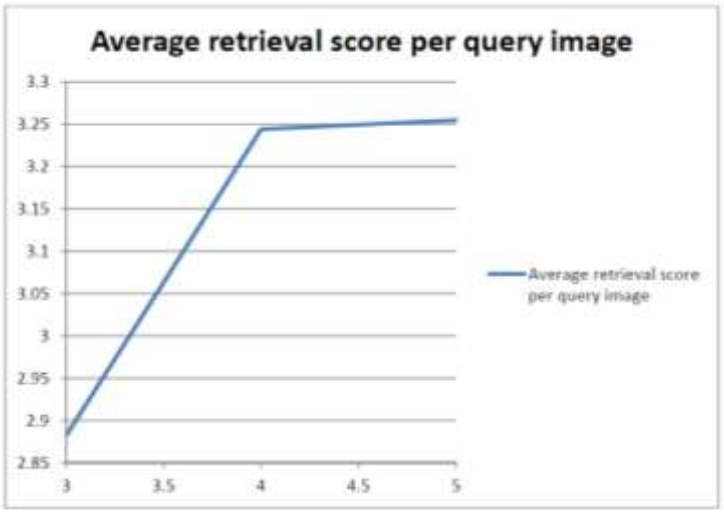
The table above demonstrates that the number of key points detected in the sample image does indeed increase as the number of scale levels per octave get greater. The question that arises here is whether the additional regions of interest which have been detected and described equate to structures in the picture that are significant to the matching problem and which therefore help to discriminate one picture from another. In other words, the additional key points must help

to describe image content that is significant and which can be used to provide a better match between the query picture and each candidate image in the collection. Once again, accuracy results are shown in the following graph and these have been recorded as the number of levels per octave is increased with each indexing step. In order to ensure that the results here are accurate, the query image in each case has been analysed with the same number of octaves and the same number of octave levels as has been applied to the collection in each case. Again, the default number of octave layers which are processed in an image by the SURF algorithm is four.



Average retrieval accuracy on a subset of the UK Bench corpus as the number of octave layers in the SURF algorithm is increased

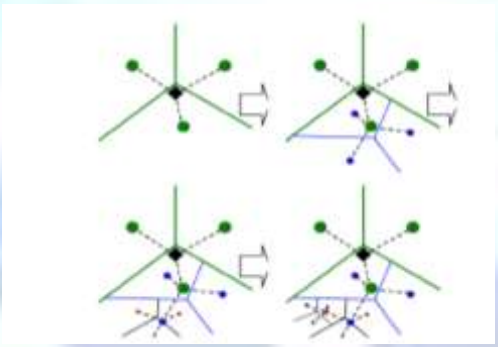
The graph in the Figure above demonstrates that retrieval accuracy does indeed increase as the number of octave layers which is used to prime the SURF algorithm is increased. The rate of the improvement here is telegraphed by the scale of the graph but it is most noticeable in the difference in accuracy which can be achieved by increasing the number of octave layers from three to the default number of four. A very small performance boost is achieved by increasing this number beyond this point, such that four octave layers gives an average accuracy figure of 81.1% and five results in an improvement to 81.3%. Although the improvement is modest, it is suggested that any gain which can be made in retrieval accuracy through the parameterisation of the SURF algorithm must be welcomed. Once again, there will be a computational overhead associated with increasing the number of octave layers which are considered because the feature detection algorithm must process the image data with an increased number of box filters in order to derive blob response results here. The same boost in performance can be viewed from the perspective of the average score which is attained by each query image in the one thousand and twenty image subset of the UK Bench corpus which demonstrated the best accuracy results in previous experiments.



Average retrieval score per query image in the UK Bench subset with increasing numbers of scale levels per octave of scale space

Once again, a marked improvement can be seen here between the scenario where the SURF algorithm is primed with three scale levels per octave and that where it is run with the default of four levels. The score is 3.24 per query with four scale levels per octave and 3.25 with five scale levels per octave. The final set of parameters which can be considered here are the branching factor and the number of levels which are used to create the scalable vocabulary tree which defines and builds the visual vocabulary. In general, increasing the number of key points which are available at the quantisation stage improves the performance of the system because there is more material to classify by centroid here. This next set of experiments acknowledges that, to some extent, improvements which can be made in this manner are marginal because the process of quantisation which is an integral part of the bag-of-words model necessarily loses information in order to be able to represent the image collection with a finite number of words that are diagnostic of the content of the pictures. Changing the indexing parameters directly is the first point at which the way that the visual vocabulary is created is changed and the results of these alterations can be assessed.

The DBow2 library which has been used to implement indexing functionality in this report follows a hierarchical process for visual vocabulary creation that was first proposed by Henrik Stewénius and David Nistér. The vocabulary tree defines a hierarchical quantisation of feature descriptors that is built by hierarchical k-means clustering. Here, instead of the value k defining the final number of clusters or quantisation cells, it instead defines an attribute called the branching factor of the tree. This equates to the number of children that will be created from each node in the tree. First, an initial k-means clustering process is run on the image collection data and k cluster centres, or centroids, are defined during this step. The resulting vocabulary is thus partitioned into k groups where each group consists of the descriptor vectors which are closest to a particular cluster centre. The same process is then recursively applied to each group of descriptor vectors such that quantisation cells are recursively defined by splitting each cell into k new parts. The tree is determined level by level, up to a maximum number of levels which can be specified in an input parameter during the indexing step. Each division of a cell into k parts is only defined by the distribution of the descriptor vectors that belong to the parent quantisation cell. In the final analysis, this process presents two additional parameters which can be altered to affect the way in which the visual vocabulary, or the bag-of-words, is created and expressed.

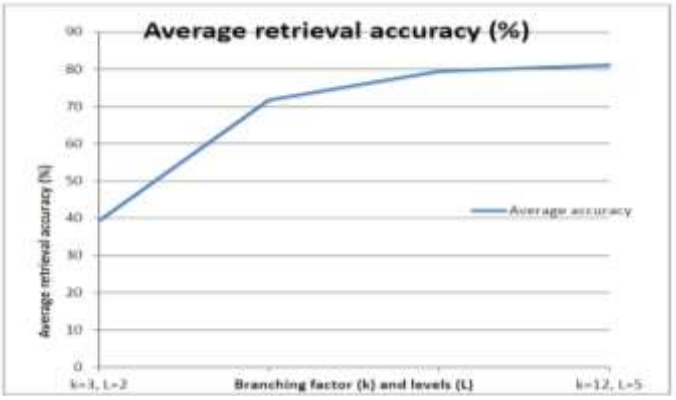


An illustration of the creation of a vocabulary tree with a branching factor of three and an arbitrary number of levels. In reality, this number of levels is constrained by an integer parameter.

In the above image, the green circles represent cluster centres and the blue and green Voronoi regions are quantisation bins around a particular visual word. In order to find the correct centroid for each candidate descriptor vector, each vector is simply propagated down the tree. This is achieved by comparing the candidate feature descriptor to the k candidate cluster centres that are represented by the children in the tree and choosing the closest one. The branching factor directly defines the number of cluster centres which will be created on each level of the vocabulary tree and the levels parameter affects the number of child quantisation steps which are performed around each existing centroid. This indicates that increasing the branching factor and the number of levels will increase the size of the visual vocabulary that is generated for a particular image collection and that; conversely, reducing these integer parameters will result in a smaller vocabulary from which a bag-of-words representation of a novel query image can be created.

It has previously been stated that the maximum number of branching levels which can be supported on the hardware platform that was available for the creation of this report is twelve. The maximum number of levels per branch that is feasible here is five. Increasing the size of these parameters beyond this point produced memory shortages and resultant errors during the indexing step. Nevertheless, the maximum supported number of branches and levels here generated a visual vocabulary of over two hundred and eighty thousand words over the full collection of ten thousand two hundred images in the UK Bench corpus. The next graph which is presented here shows the effect of reducing the branching factor and the number of tree levels from the maximum supported numbers. In each case, the resultant visual vocabulary becomes smaller and the effect of this on retrieval accuracy and the average retrieval score that is achieved by each query image is analysed. Once again, the results have been compiled from processing the first one thousand

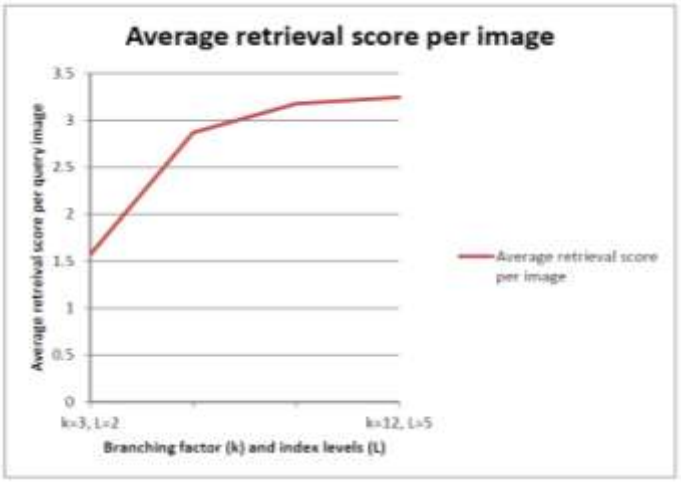
and twenty images in the UK Bench collection which give rise to the best results for accuracy and score per image in previous experiments.



Average retrieval accuracy on a subset of the UK Bench corpus with different degrees of index branching and index levels

A significant drop off in retrieval accuracy is evidenced with smaller values of k and L, as would be expected from the preceding discussion. This means that the feature space of the collection is quantised into just nine cluster centres. It is obvious that this scale of processing does not provide the discriminative power which is required for the bag-of-words system to function acceptably over a corpus of thousands of images. A smaller visual vocabulary is significantly quicker to compute than a more complete version with more centroids and a balance needs to be struck here between retrieval accuracy and the capabilities of the host computer system to generate larger vocabulary trees efficiently. The values of k and L which have been used here to generate the graph have not been chosen randomly. They have been taken from various sources in the research literature as natural stepping increments for the vocabulary generation process. In general, a larger vocabulary is better both because it is more discriminative and because the process of creating it is a post-query or offline step which only needs to be completed once.

The size of the vocabulary tree which is used does not have a significant effect upon query times across the database when an inverted index structure is used and this has been demonstrated by the timing results that have been obtained in previous experiments. The time required to query a single image against the whole ten thousand two hundred image UK Bench corpus remains sub-second. The authors of the corpus have achieved slightly better average retrieval score results than have been reproduced in this report and it is likely that they have been able to produce a larger visual vocabulary to represent the collection than was supported by the hardware available in this case. The memory requirement for indexing and particularly for the hierarchical quantisation step is significant and the amount of resource which is available here is the main limiting factor in the creation of large bag-of-words image representations. Nevertheless, the same drop off in retrieval accuracy is evidenced in terms of the average retrieval score per image, which is shown in the next graph.



Average retrieval score per query image on a subset of the UK Bench corpus with different degrees of index branching and index levels

3.4 The effect of Query Image Resolution and Quality

The final research question which is worth examining in relation to the bag-of-words image retrieval system is what effect the resolution and quality of query pictures has on feature detection and the levels of accuracy which can be achieved. It may be, for example, that poor optical resolution will produce fewer key points in a picture because there is less granularity about the intensity information which is encoded. On the other hand, image compression can itself produce artefacts which are introduced by the storage procedure and these features may be misidentified by the analysis algorithms as significant regions of interest within a picture which could affect the matching performance of the bag-of-words system. In order to analyse the effect of image quality and resolution on the image matching process, an experiment has been designed which involves taking a query picture from the Open University art collection which was captured on a mobile device. The image will be progressively compressed to different quality levels using the JPEG algorithm. First of all, the number of key points which are identified in the picture at each level of compression will be recorded and secondly, the query will be manually tested for matching performance against the corpus of pictures to establish whether any discernible effect on the accuracy of retrieval is demonstrated.

The sample query image from the OU Art corpus that has been used to test the effect of query image quality on key point detection and retrieval accuracy

The image that has been used as a test query here is shown in the figure above. In normal circumstances, this picture produces the correct match at search rank position one in the results list. It also furnishes three correct alternative images from the corpus which all feature different viewpoints of the same object. The system has been configured here to produce four search results in order of relevance for each given query and so this means that one of the result images which are typically returned is not correct. The incorrect picture is completely incorrect in terms of the query with a different object being shown to that in the interrogation request. There is in fact no apparent similarity between the given query and the incorrect result image in the group of four although this error comes in at position four on the ranking and so it does not affect overall retrieval accuracy under the scheme that has been used to measure performance with the OU Art corpus. In the initial instance, the number of key points which are detected in the query image can be examined as the quality of the picture is reduced in ten percent increments by increasing the level of JPEG compression at each stage.

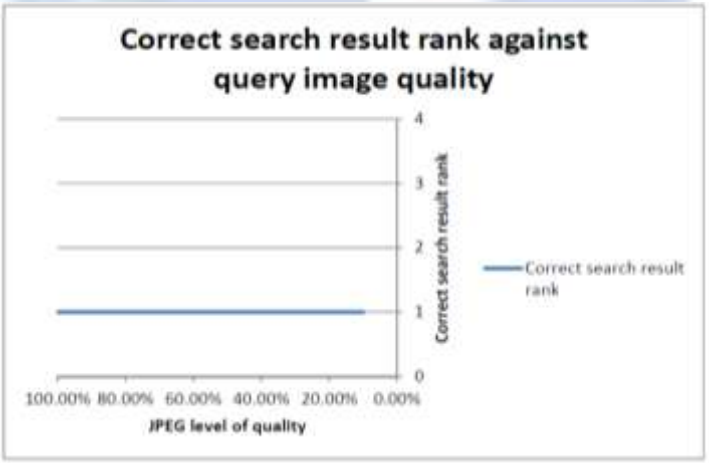
JPEG Quality level	Number of keypoints detected
100%	971
90%	962
80%	997
70%	987
60%	957
50%	977
40%	973
30%	939
20%	944
10%	954

JPEG Quality level	Query file size (kilobytes)
100%	89
90%	57.3
80%	43.6
70%	35.6
60%	29.8
50%	26.2
40%	22.7
30%	19.2
20%	14.7
10%	9.2

The results which are shown in the above two tables are interesting. Theydemonstrate that, whilst the file size of the query image decreases progressively withhigher levels of compression, the number of key points which is detected in thepicture does not reduce in a linear fashion. Indeed, the number of key points whichare identified actually increases at stages as the quality level of the JPEG file isreduced and the file size itself gets smaller. One explanation for this is that thecompression mechanism creates artefacts and structures within the image which areused as the basis for regions of interest that can be detected and described by theSURF algorithm. In fact, the overall decrease in the number of key points betweenthe full quality image and the very lowest quality rendering is only some seventeenpoints. This figure hardly seems significant in terms of the overall number ofregions which have been detected and described in the picture. In order toemphasise the effect that compression has on the quality of the query picture, thelowest quality version of the image is shown in the following figure.

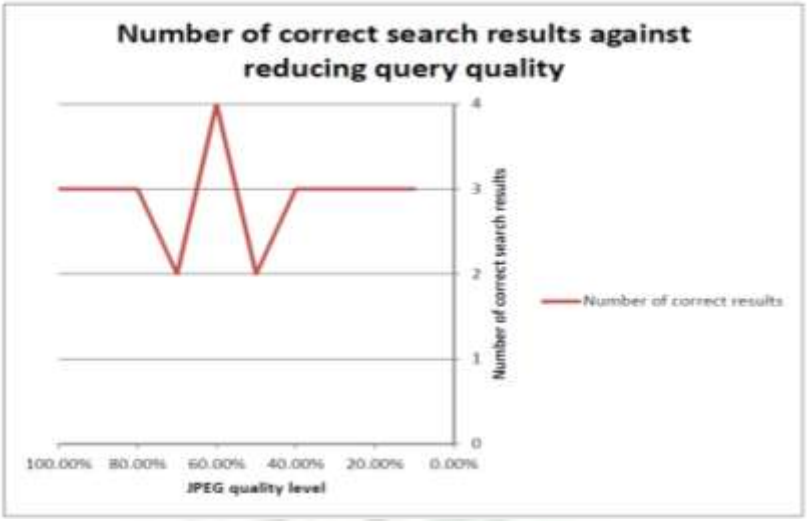
The same query image shown with the highest level of compression applied

It is quite clear from the above picture that the compression process has discarded detail and information from the query. Nevertheless, the fundamental structure and object of the photograph remain clear and it may be that the lossy algorithm has in fact simplified the image to the extent that image features which were previously detected on detailed regions of the picture are progressively concentrated on more significant changes in the gradient and intensity data. This may lead to a better diagnosis of the matching images in the collection because spurious regions of interest which are not of critical importance to the matching problem are gradually eliminated. In any event, the following graph highlights the retrieval performance which is seen when comparing the progressively compressed picture against images in the collection.



The rank of the first correct search match in search results against a progressively compressed query image

These results show that compressing the query image and thereby reducing the quality and detail of the picture do not affect retrieval performance to the extent that the image is incorrectly matched to a candidate picture in the collection. The correct search result is always returned at ranking position one within the four images that are returned in response to the query. Indeed, it was anticipated that the correct image might drop down the search result rankings as the level of compression on the query increased but this has been proven to be an unfounded expectation. This is good news for real world content-based image retrieval platforms since it demonstrates that the bag-of-words approach works well with query pictures that are compressed to a very high degree. The full picture of how compression and reduced image quality affect the matching process is somewhat more complicated, however, as the following figure illustrates.



The number of correct search results returned against increasing levels of JPEG compression

The above figure demonstrates that the number of correct search results which are returned in response to a progressively compressed query picture drops off at certain levels of quality but actually increases with a specific degree of compression. The picture which has been used as a query here happens to match correctly to four candidate pictures in the collection and, under normal circumstances with a pristine interrogation request, the bag-of-words system returns three of these possibilities in the top four search results. However, when the quality of the query image is reduced by applying JPEG compression to the sixty percent level, the system manages to correctly match all four of the potential candidates to the interrogation request. This result suggests that there is a level of compression available which actually simplifies the picture so that spurious features are discarded and the remaining regions of interest can be centred on portions of the image which are critical to there solution of the matching problem. It follows that there is a balance to be struck between reducing the quality of the query so that new and misleading artefacts are created in the image and simplifying the picture so that more significant features can be prioritised in the matching process. Some researchers at the Open University have indeed found that pre-processing steps such as reducing the physical size of thequery image can lead to better matching performance because the level of detail in the picture is reduced. A more detailed set of experiments would need to be conducted in order to determine whether the improvement in retrieval accuracy which is evidenced here at sixty percent quality settings on the query image could be replicated more broadlyand reliably across different pictures and collections of images.

CONCLUSIONS AND FUTURE WORK

This paper has been concerned with the development and testing of content-based image search and retrieval platforms which scale effectively to work with large numbers of pictures. The idea is to provide online search functionality so that a query is expressed in terms of an image and the system then returns results which are also pictures and which are ranked according to how similar they are to the interrogation request. The content of the image itself becomes the information upon which documents are matched in this scenario. The concept of similarity here is necessarily defined according to human perception of the appropriateness of a particular search result in the context of a query picture. However, the algorithms which provide the matching facility work by identifying features of different images which are probably not directly obvious to a human when they arrive at a conclusion about whether one picture matches another one or not.

A prototype search and retrieval system which uses the bag-of-words approach is presented. Indexing times for a growing collection of pictures are examined together with thetime taken to query a single picture against the corpus as the scale of that corpus rows. The two main collections of pictures which are used for evaluation purposes are the Open University and the UK Bench collection, respectively. Practical ways to measure the performance of different systems on these collections of images are then considered and a system for the generation and comparison of appropriate metrics is provided.

The key contribution of this research work is to establish an automated system for measuring retrieval accuracy against the large scale UK Bench corpus which consists of ten thousand two hundred pictures. This collection is broken down into subsets and each subset is composed of groups of four images which all feature the same object from different viewpoints and perspectives. Therefore a simple algorithm can be developed which checks for query efficiency and

accuracy against these groups without the need for human intervention or for manual consideration of the similarity of query images to those results that are returned from the collection in each case.

The findings of this report are that image search and retrieval techniques show great promise in that they provide accurate information in response to the majority of queries. The semantic gap problem is clearly an area for further work because, when the search results do break down and inaccurate or irrelevant material is provided, there is little that a user can do to reformulate their query or to ensure that similar unwanted information does not appear in response to subsequent interrogation requests. The error rates that have been demonstrated with the bag-of-words system are still quite high when these are measured across the totality of a response set but the impact of this flaw is ameliorated by the fact that a highly appropriate response is given in position one of the result set in the majority of instances. It has also been shown that not all image search requests provide the same level of difficulty. Pictures can be taken and composed in such a way that the object of interest which the user is seeking information about is significant and central in the frame of the picture. This helps to ensure that features are generated from portions of the image which are important to the matching problem in a given scenario.

Other images which feature small objects or which show items from a long way away are a much harder matching proposition and they give rise to most of the errors that have been seen with the UK Bench corpus of pictures. In general, however, the invariance of local image features works to good effect and the robust level of matching which canbe seen in commercial products such as Google Goggles is backed up in the results of this research. Indeed, the performance of the Goggles platform often outstrips what has been possible here because the technology uses various additional techniques in order toboost retrieval performance. The integration of pattern recognition and text analysis capabilities often provides a fall back in this scenario where objects that would not otherwise be matched correctly at least lead to some useful information in the search results because details have been correctly identified and mapped to textual information that is of interest to the user. Obviously the scope of the Goggles platform is outside the remit and capabilities of this study but the work here does suggest that research platforms can expect to achieve similar levels of performance as seen in technologies such as Goggles and SnapTell without the same level of financial or resource investment that is evident there. It is suggested finally that this research introduces originality in its findings because it is impossible to find another study which encompasses both an examination of the theory behind content based image retrieval and a detailed set of benchmarks over publicly available image corpora.

The authors of the UK Bench collection themselves provide some metrics for performance with the collection but it is not as comprehensive or considered as the results that have been obtained here. Content based image retrieval is an effective and scalable technology which must answer philosophical problems to do with the gap between human visual perception and a computer model of image matches urgently. Further technical developments in the field must not just be based upon algorithms that cannot be justified and have not been designed with regard to accepted standards for their mathematical structure and operation.

References

[1]. Datta, R., Joshi, D., Li,J., &Wang, J. (2008). Image retrieval: Ideas, influences, and trends of thenewage. ACM Computing Surveys (CSUR),40(2), 5.
[2]. Li,J., &W ang, J. (2008). Real-time computerized annotation of pictures. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30(6), 985-1002.
[3]. More Thoughts on Image Retrieval.(2009, 05 08).Retrieved from The Noisy Channel: http://thenoisychannel .com /2009 /05 /08 /more-thoughts-on-image-retrieval/
[4]. Hirata, K., & Kato, T. (1992). Query by visual example. Advances in Database Technology—EDBT'92, (pp. 56-71)
[5]. [5]. Eakins, J., & Graham, M. (1999). Content-based image retrieval. Programme of Joint Information Systems Committee on Technology Applications. JTAP.
[6]. Pavlidis, T. (2008). Limitations of content-based image retrieval. Invited Plenary Talk at the 19th Internat. Conf. on Pattern Recognition.
[7]. Enser, P., & Sandom, C.(2003). Towards a com prehensive survey of the semantic gap in visual image retrieval.Image and Video Retrieval, 163-168.
[8]. [8]. Chiu, C.-Y., Lin, H.-C.,&Yang, S.-N. (2003). Learning Human Perceptual Concepts inaFuzzyCBIRSystem.Proceedingsofthe5th International Conference on Computational Intelligence and Multimedia Applications Washington, DC,USA: IEEE Computer Society.
[9]. Thiagarajan, R., Manjunath, G., & Stumptner, M.(2008). Computing semantic similarity using on tologies.ISWC08, the International Semantic Web Conference (ISWC).
[10]. Colombino,T., Martin, D., Grasso, A.,&Marchesotti, L. (2010).A Reformulation of the Semantic Gap Problem in Content-Based Image Retrieval Scenarios.I nt. Conf.on the Design of Cooperative Systems.
[11]. Wang, H., Mohamad, D.,& Ismail, N. (2010). Semantic gap in CBIR: Automatic objects spatial relationships semantic extraction and representation. International Journal of Image Processing (IJIP),4(3), 192.
[12]. Wang, C., Zhang, L., & Zhang, H. (2008). Learning to reduce the semantic gap in web image retrieval and annotation. Proceedings of the 31st annual international ACMSIGIR conference on Research and development in information retrieval, (pp. 355-362).

[13]. Liu, Y., Zhang, D., Lu, G., &Ma, W. (2007). A survey of content-based image retrievalwithhigh-level semantics.PatternRecognition,40(1), 262-282.
[14]. Hu, R., Ruger, S., Song, D., Liu, H., & Huang, Z. (2008). Dissimilarity measures for content-based image retrieval. Multimedia and Expo, 2008 IEEE International Conference on, (pp. 1365-1368).
[15]. Liu, H., Song, D., Rueger, S., Hu, R., &Uren, V.(2008). Comparing dissimilarity measures for content-based image retrieval. Information Retrieval Technology, 44-50.
[16]. Deselaers, T., Keysers, D., & Ney, H. (2008, #apr#). Features for image retrieval: an experimental comparison. Inf. Retr., 11(2), 77-107. Retrieved from http://dx.doi.org/10.1007/s10791-007-9039-3
[17]. Little, S., Brown, E., & Rueger, S. (2011). Multimedia: information representation and access.
[18]. Lowe, D. (1999). Object recognition from local scale-invariant features. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, 2, pp. 1150-1157.
[19]. Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.
[20]. Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. Computer Vision-ECCV 2006, 404-417.
[21]. Bauer, J., Sunderhauf, N., & Protzel, P. (2007). Comparing several implementations of two recently published feature detectors. Proc. of the International Conference on Intelligent and Autonomous Systems.
[22]. Tola, E., Lepetit, V., &Fua, P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(5), 815-830.
[23]. Delponte, E., Arnaud, E., Odone, F., &Verri, A. (2006). Analysis on a local approach to 3d object recognition. Pattern Recognition, 253-262.
[24]. Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2161-2168.
[25]. Gálvez-López, D., & Tardós, J. D. (2012). Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Transactions on Robotics, Volume 28, Number 5, 1188-1197.