An Adaptive Technique for Content-Based Image Retrieval

Jana Urban, Joemon M. Jose and Cornelis J. van Rijsbergen
({jana,jj,keith}@dcs.gla.ac.uk)
Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK

Abstract. We discuss an adaptive approach towards Content-Based Image Retrieval. It is based on the Ostensive Model of developing information needs—a special kind of relevance feedback model that learns from implicit user feedback and adds a temporal notion to relevance. The ostensive approach supports content-assisted browsing through visualising the interaction by adding user-selected images to a browsing path, which ends with a set of system recommendations. The suggestions are based on an adaptive query learning scheme, in which the query is learnt from previously selected images. Our approach is an adaptation of the original Ostensive Model based on textual features only, to include content-based features to characterise images. In the proposed scheme textual and colour features are combined using the Dempster-Shafer theory of evidence combination.

Results from a user-centred, work-task oriented evaluation show that the ostensive interface is preferred over a traditional interface with manual query facilities. This is due to its ability to adapt to the user's need, its intuitiveness and the fluid way in which it operates. Studying and comparing the nature of the underlying information need, it emerges that our approach elicits changes in the user's need based on the interaction, and is successful in adapting the retrieval to match the changes. In addition, a preliminary study of the retrieval performance of the ostensive relevance feedback scheme shows that it can outperform a standard relevance feedback strategy in terms of image recall in category search.

Keywords: content-based image retrieval, adaptive retrieval, ostensive relevance, relevance feedback, user evaluation

Abbreviations: CBIR – Content-based Image Retrieval; RF – Relevance Feedback; OM – Ostensive Model

1. Introduction

The *semantic gap* has become a buzzword in Content-based Image Retrieval (CBIR) research. It refers to the gap between low-level image features and high-level semantic concepts. Despite considerable research effort in this field over the last decade, there has not been any significant success for generic applications. Today, the research community has accepted the fact that it will probably be impossible to retrieve images by semantic content for several years. Instead, people have started to exploit relevance feedback techniques. Relevance feedback is regarded as an invaluable tool to improve CBIR effectiveness, not only because it provides a way to embrace the individuality of users, but it is indispensable in bridging the semantic gap.

The semantic gap has further implications on the query formulation process. Since low-level features do not directly reflect the user's high-level perception of the image content, the query formulation process is even more difficult than in text retrieval systems. Moreover, the underlying search need is dynamic and evolving in the course of a search session. Most often, image searching is explorative in nature, where searchers initiate a session and learn as they interact with the system. However, current CBIR systems fail to deal with the dynamic nature of search needs.

In this work, we introduce an adaptive retrieval system, which places particular emphasis on the "human in the loop". In the proposed system the retrieval process is iterative, updating the system's knowledge of the user's information need based on the user's implicit feedback. To this end, it incorporates an adaptive image retrieval technique based on the *Ostensive Model (OM) of developing information needs* [3]. In the underlying interaction model the user builds up a browsing tree of interesting images by choosing one image from a recommended set to be appended to the browsing path in each iteration. The system's recommendations are based on a query constructed from the current path of images. For the query, each image in the path is considered relevant, but the degree of relevance is dependent on age: it decreases over time when new images are appended. In this way, the OM is a special kind of relevance feedback model, in which a query is refined by the user implicitly selecting images for feedback. It recognises and addresses the issue of *dynamic nature of information needs*, and has the advantage of allowing for an intuitive and user-centred search process.

In order to evaluate the effectiveness of the ostensive relevance approach, we built three systems: a baseline system with manual query facilities and two variances based on the OM. Results of a user-centred, work task-oriented evaluation show that the ostensive browsing interfaces are generally preferred over the comparative system. Its strengths are considered to lie in its ability to adapt to the user's need, and its very intuitive and fluid way of operation.

In addition, the retrieval performance of the underlying technique is evaluated in a simulated environment assuming a user conducting category search. In comparison to a traditional relevance feedback technique as baseline, it shows that the query learning scheme based on the OM can outperform the baseline strategy in terms of the total number of images belonging to a specific category found.

The main contributions of this paper are three-fold. First, we show how the OM can be adapted to both textual and content-based features with the proposed query learning scheme. Second, we provide results from an extensive user evaluation of CBIR interfaces. Last but not least, we highlight the merits of the ostensive relevance feedback strategy in terms of retrieval performance and point out future directions for a continuative evaluation.

The remainder of the paper is organised as follows. Section 2 provides an account of the motivations for our approach: the intrinsic difficulties of the query formulation process and a review of current approaches to relevance feedback. In particular, these are compared to the notion of ostensive relevance. In Sections 3 & 4 we introduce the systems we used for the evaluation and describe the proposed adaptive query learning scheme. The experimental methodology is detailed in Section 5, followed by a review of our experimental results. The main findings of the user study are summarised and discussed in Section 7, while Section 8 adds results of a quantitative evaluation of the underlying query learning scheme. Finally, we conclude and provide a brief outlook to the future.

2. Content-based Image Retrieval

Every information seeking process is necessarily initiated by an information need on the user's side. Therefore, the success of a retrieval system depends largely on its ability to allow the user to communicate this need. For the user, the query formulation often poses a significant hurdle due to manifold reasons, which will be discussed briefly in this section. A popular means to overcome these problems is *relevance feedback*. This section therefore also covers the idea behind relevance feedback and discusses some examples in the CBIR literature.

2.1. Query Formulation Problem

One of the major issues in information searching is the problems associated with initiating a good query. However, it has been well accepted that searchers find it hard to generate a query due to the following reasons [21]. Firstly, searchers do not know how the documents are represented. This is especially true for CBIR systems due to the low-level representation of content-based features. It is substantially difficult to formulate a query in terms of the system's internal representation of the documents, which is often composed of a collection of low-level features in the pixel domain.

Secondly, the underlying information need itself is typically vague (*"I don't know what I'm looking for, but I'll know when I find it"* [28]).

Due to the uncertainty about what information is available or about the actual need itself, a search process usually starts with an *explorative phase*, in which the user tries the system and its options, and tests what kind of information is returned. Through exposure to new objects, the user's context is changing and their knowledge state is developing, triggering changes in the user's need [10, 8]. This often leads the user to reformulate the initial query either to make it more precise after having gained some knowledge about the collection make-up, or to steer it in different directions after having seen other interesting documents, or a combination of both.

2.2. Relevance Feedback

In order to alleviate the query formulation problem, a popular approach is to incorporate relevance feedback into the retrieval system. Relevance feedback is a commonly accepted means to improve retrieval effectiveness and has been studied extensively (e.g. [19, 21]). In IR systems incorporating relevance feedback, the search process is initiated with a user-supplied query, returning a small number of documents to the user. The user is then given the possibility of indicating which of the returned documents are actually useful (relevant). Based upon those user relevance judgments the original query is automatically reformulated. The new query is again matched against the documents in the collection, returning an improved set of documents. This process can continue until the user's information need is satisfied. As a consequence, a retrieval system based on relevance feedback is inherently interactive.

As mentioned earlier, relevance feedback is a tool to kill two birds with one stone. It is used to bridge the *semantic gap* by avoiding or helping the query formulation process, while at the same time it naturally provides a way to embrace the individuality of users. The user's judgement of relevance is naturally based on their current context, their preferences, and also their way of judging the semantic content of the images (e.g. [23, 7]). Initially, the system uses the low-level image features as a quick way to 'estimate' the relevance values of the images. By prompting the user for relevance feedback, this rough estimation can be improved to steer the results in the direction the user has in mind. A comprehensive study of existing relevance feedback techniques in image retrieval can be found in [35].

2.2.1. *Assumptions*

Different methods have been adopted on the basis of often diverging assumptions. One major variance is *what* actually is fed back to the system. Often, binary feedback for positive and negative examples is used (e.g., [29]), some additionally associate a 'degree of (ir)relevance' (e.g., [18]), and others interpret the feedback only as a 'comparative judgment' (e.g., [6]). Depending on the assumptions taken in this respect, the resulting systems can be distinguished further: While positive feedback has been used for feature selection (e.g., [17]) or feature relevance weighting (e.g., [18, 11]), using both positive and negative feedback gives rise to treating the retrieval process as a classification or learning problem. Many systems now strike the latter path, transferring methods previously employed mainly in the field of artificial intelligence (e.g., [30, 34, 29]). However, they are hindered by one major obstacle, namely the *small sample issue* [35]. The user feedback in each iteration only gives a tiny number of training samples relative to the high dimension of the feature space and the possible number of classes for general multimedia data. Consequently the results are unreliable on their own, requiring a lot of extra effort, e.g. an off-line training phase following the on-line search as employed in [34] to arrive at meaningful results. This is often undesirable, since it militates against the real-time requirement of relevance feedback. The main advantage of relevance feedback, namely that it allows real-time learning from user interaction to improve the system's performance during one search session, is thus undermined.

A further characteristic of existing systems is *how* they gain information about the user's judgment of relevance. One can distinguish between two distinct approaches: *explicit* and *implicit* relevance feedback. Explicit relevance feedback, which is assumed in most current systems (e.g., [6, 18, 30]), asks the user to explicitly state whether a returned document is relevant or not. Therefore, the user interface has to provide for facilities to input this judgment by the user. This (additional) task is often considered as a burden to the user, since it is difficult for most users to assess the degree of relevance of one document in terms of a numeric value [33], which presumes considerable knowledge of the retrieval environment. Even though it might be much easier to determine whether an image is actually relevant to the user compared to formulating a good query, it still requires often considerable cognitive effort from the user to communicate this relevance assessment to the system. For this reason, a less-distracting possibility to gain relevance feedback is implicitly from the users, simply by observing their interaction with the system [32].

Another assumption underlying nearly all current relevance feedback techniques is that a user's information need is static and there is no provision for updating user's judgements. Especially those techniques that attempt to classify or separate the document space into relevant and non-relevant, explicitly rely on the assumption that—within one search session—all documents are either relevant or not regarding the user's information need. In other words all documents are assumed of having constant relevance values. However, this is a rather simplifying view of the real-world. Not only are the user's *actions* time-dependent—resulting in giving inconsistent feedback, but even more importantly, the user's *goals* are also time-dependent and might change either gradually or quite abruptly. The trigger for such changes is most often a result of having come across something interesting that they have not even considered at the beginning of the search. For this reason the system proposed here is based on the *Ostensive Model*, which captures *"the intentionality of an information need that is assumed to be developing during the searching session"* [2]. The details of the model will be described in Section 2.4.

In order to avoid asking the user for explicit relevance feedback, the approach taken here is the one of interpreting a user's selection of one document over others as an indication that this document is more relevant. Due to this 'ostensive' approach, and in fact the Ostensive Model underlying this system, only positive examples are there to work with. The positive feedback is used for query learning, in which the query itself is learnt and subject to adaptation. On the basis of the documents selected, the system creates a new query consisting of a combination of these documents' features. This query adapts in every iteration of the retrieval process.

2.3. Query Learning Approaches

To relieve the user from the query formulation problem, a method that is able to "guess" or "learn" the user's desires purely from a set of examples is believed to be very advantageous. Algorithms for CBIR that rely on query refinement as a way of incorporating relevance feedback have attracted a lot of interest (e.g., [11, 18, 20]). They are based on the early work on *query shifting* in the text retrieval domain [19]. Query shifting is the prevalent technique of adapting an initial query, which aims at moving the query toward the region of the feature space containing the set of relevant documents and away from the region of the set of non-relevant documents [11, 18, 20]. The underlying assumption is that the user has an *ideal* query in mind, and the system's task is to find this ideal query. Often query refinement methods are used in combination with *feature re-weighting*, which is based on a weighted similarity measure where relevance feedback is used to update the weights associated with each feature in order to model the user's need [11, 18, 20].

The approach proposed in this paper is a form of *dynamic* query learning, combining both query shifting and feature re-weighting techniques. Unlike other query learning methods, which lack the ability to adjust the degree of relevance over time, the emphasis in our approach lies on dynamically adapting relevance values. The temporal dimension to the notion of relevance introduced in the Ostensive Model is exploited to achieve this different view of adaptive query. The details of the model will be described in the following.

2.4. Ostensive Relevance

The Ostensive Model (OM) of developing information needs was initially proposed by Campbell and van Rijsbergen [4]. It combines the two complementary approaches to information seeking: query-based and browse-based. It supports a query-less interface, in which the user's indication of interest in an object—by pointing at it—is interpreted as evidence for it being relevant to their current information need. Therefore, it allows direct searching without the need of formally describing the information need. The query is automatically evolved from a path of documents selected in the course of one search session.

By accepting that the user's need is dynamically changing during a search session, the OM adds a temporal dimension to the notion of relevance. A recently selected object is regarded more indicative to the current information need than a previously selected one. So, in this sense, the degree to which a document is considered relevant is continuously updated to reflect the changing context. The definition of Ostensive Relevance summarises the main points [3]:

> The **Ostensive Relevance** of an information object is the degree to which evidence from the object is representative/indicative of the current information need.

The interaction with an Ostensive Browser follows an intuitive scheme. Here the user starts with one example document as the query, and as a result is presented with a new set of candidate documents (top ranking documents according to the similarity measure used). As a next step, the user—through selecting one of the returned documents—updates the query, which now consists of the original document and the selected document of the set of returned candidates. After a couple of iterations, the query is based on a path of documents. Similarly to the Path Model described in [5] for activity-centred information access, emphasis is set on the user's activity and the context, rather than the predefined internal representation of the data. A path represents the user's motion through information, and taken as a whole is used to build up a representation of the instantaneous information need.

The weight of how much each document along the path contributes to the next query can be chosen with different objectives in mind. The weighting schemes are referred to as *ostensive profiles*, and reflect

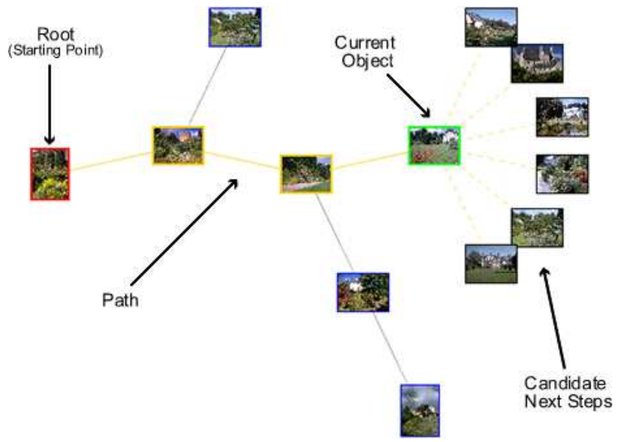


Figure 1. The ostensive path

how relevance (or uncertainty) changes with age (age being interpreted as the order of selection or the position along the path). With the previously elaborated considerations in mind, the most plausible profile supports uncertainty increasing with age. The further back in time one document has been selected during the retrieval process, the more uncertainty is associated with it that it actually reflects the user's information need, or in other words, the less relevant it is considered for the query. This profile is also the one favoured by the original definition of Ostensive Relevance. For a comparative evaluation of different profiles and their interpretations please refer to [2].

Since the whole path is visible to the users, they can jump back to a previous object along the path if they get the feeling that they are stuck or moving in the wrong direction. From there a new path can be explored, starting from the original object (the root) and the newly selected object. The resulting paths form a tree-like structure, originating from one root and branching at various objects (see Fig. 1).

The OM thus captures the developing information need of the user during a search process, and incorporates the uncertainty, which necessarily exists due to the imprecise awareness of one's own information need and the difficulties of expressing it.

In its original conception the OM was integrated with the Binary Probabilistic Model (BPM) of IR to create an operational retrieval model [3]. This was possible, since the images that were used in the implementation of the OM were represented by a set of index terms. However, if one takes into account content-based features to index images, the interpretation of the BPM becomes rather difficult. In the BPM, relevance scores are based on estimating or calculating the probabilities that, if the document is relevant (or non-relevant respectively), a particular feature will be observed. In other words, the probability is assessed depending on whether some chosen feature is either present or absent. This interpretation was developed in the text retrieval domain, where a document can be represented by a set of index terms only. CBIR systems rely on more complex indexing features, in which it is hard to tell whether a particular feature can be observed. It is questionable whether or not content-based image features can be treated in a binary fashion, e.g. is it sensible to say the image contains the index term "green" if the colour histogram contains non-zero values for the bins referring to green? What makes matters even more complicated is the fact that most CBIR systems rely on multiple representations of image content. It becomes apparent that the interpretation of the binary probabilistic model in terms of content-based image features is rather inappropriate. For this reason, we introduce the use of adaptive queries within an operational retrieval system based on the OM.

3. The Systems

To test our ideas about adaptive query learning strategies, three prototype system have been implemented and evaluated. In this section we will describe these systems.

3.1. Features & Similarities

The systems use two distinct features: *text annotations* and *visual features*. The text feature is extracted from the keyword annotations of the images, and the visual feature is based on colour histograms representing an image's global colour distribution represented in the HSV colour space.

An image is represented by a two-dimensional feature vector, which is a term vector (text feature) and a histogram bin vector (colour feature) respectively. The term vector is weighted by the $tf \times idf$ (term frequency, inverse document frequency) weighting scheme. The similarity between documents is calculated as the combined score of the two similarity values for each feature using the Dempster-Shafer combination (see Section 4.1). In the case of text similarity, the *cosine measure* [22] is used:

$$sim(D, Q) = \frac{TV_D \cdot TV_Q}{||TV_D||\ ||TV_Q||}$$

while the visual similarity is determined by *histogram intersection* [27]:

$$sim(D,Q) = \frac{\sum_{i=1}^{l_H} min(H_D[i], H_Q[i])}{min(||H_D||, ||H_Q||)}$$

where TV stands for a document's term vector, H for its colour histogram vector, and l_H for the histogram vector length. Both similarity measures are widely used in combination with the chosen feature representation.

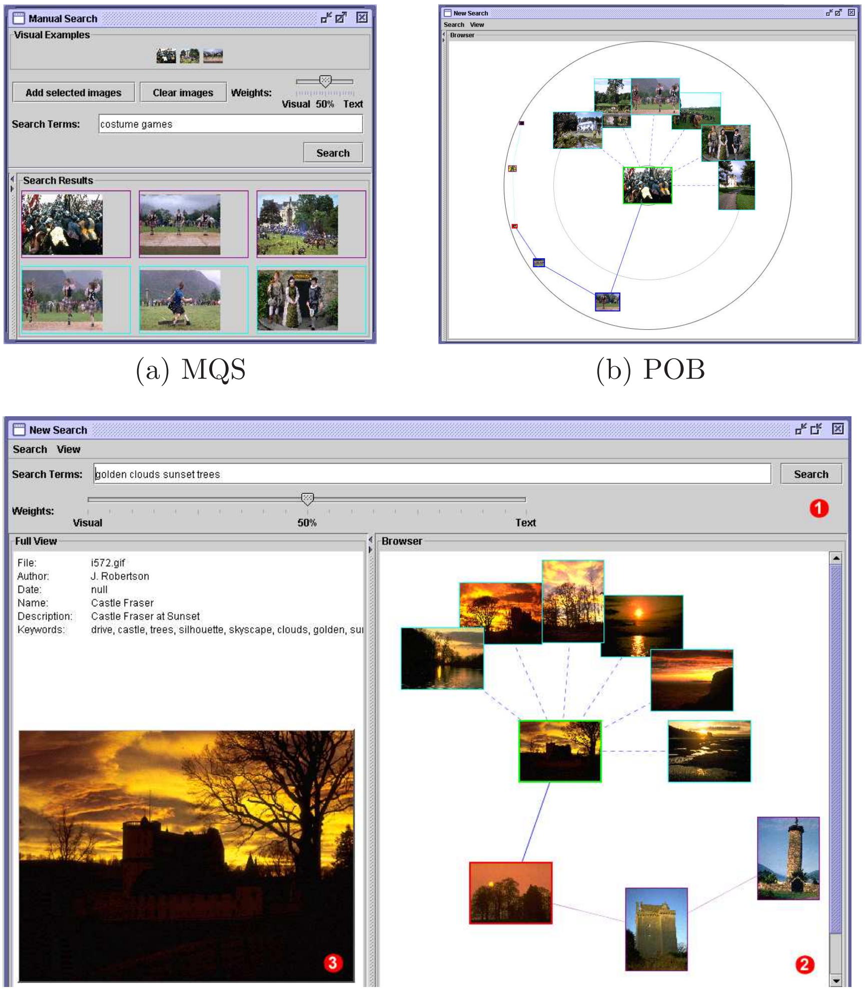
3.2. The Interfaces

3.2.1. *The Ostensive Browsers*

Two versions of the ostensive browsing approach have been implemented: one with a pure ostensive browsing scheme (Fig. 2(b)) and the other allowing explicit feedback within ostensive browsing (Fig. 2(c)). In both systems the user starts with an image in the browse panel (in Fig. 2(c)-2). The initial image is obtained in a pre-keyword search from which the user is given the opportunity to choose an image to explore further. When selecting an image, the system returns a set of most similar images as candidate images. We chose to present six images as new candidates. Of those candidates, the user clicks on the most appropriate one. At this stage, the system computes a new set of similar images based on an adapted query and presents it to the user. As in Figure 2(b) & (c), this process creates a path of images, which is represented in the interface. At any point the user can go back to previously selected images in the path and also branch off, by selecting a different candidate. The complete search session can continue to iterate between keyword search followed by browsing sessions, as long as the user is not satisfied with the retrieved images. Since the screen space is very limited the different paths are often overlapped resulting in a large degree of clutter, a fish-eye view as alternative (see Fig. 2(b)) is provided. The user can switch between these views during the search.

To view details of the image, there is the possibility of viewing a single selected image in full-size in a separate panel (in Fig. 2(c)-3). It also contains some meta-data about the document, such as the photographer, title, date, and description. In between the full-size view and the thumbnails, a quick view is shown as a popup when the user moves the mouse over a thumbnail in the browse panel.

Both browsers (Fig. 2(b-c)) attempt to adapt the query based on the user's implicit feedback, which will be described in Section 4. We provided two slightly different versions of the Ostensive Browser to allow for different levels of control. The ***Pure Ostensive Browser*** (POB) (Fig. 2(b)) does not allow for any control of feature terms or



Search Terms: costume games

Search Terms: golden clouds sunset trees

File: i572.gif
Author: J. Robertson
Date: null
Name: Castle Fraser
Description: Castle Fraser at Sunset
Keywords: drive, castle, trees, silhouette, skyscape, clouds, golden, sui

(a) MQS

(b) POB

(c) COB

Figure 2. The interfaces.

weighting between the features. The system automatically adapts the query and also the feature weights. The learning of the feature weights is achieved in a similar fashion to [18], and they will be used as trust values in Dempster-Shafer's evidence combination (see Section 4.1) to combine the similarity scores.

In addition, the interface for the ***Controlled Ostensive Browser*** (COB) provides options for selecting the features and their associated weights (in Fig. 2(c)-1). It displays the search terms the system used to obtain the currently shown candidates. The automatically selected terms (the strategy of the selection is described in Section 4), can be changed by the user and thus the current candidates are exchanged for the ones resulting from the updated query. Another aspect of control is the adjustment of the feature weights. The user can control the weights between the two features by means of a slider.

How to start the search? The problem with the ostensive search is the question of how to initiate the search, i.e. how to choose the first image that starts the path. As mentioned earlier, the current solution is to ask the user to formulate a keyword query, which returns a set of relevant images based on the textual feature. One of the returned images can then be chosen as the starting point. However, this is a rather ad-hoc approach, which again requires the user to formulate a query. We are currently investigating different solutions to this problem. One approach could be to pre-cluster the whole collection, and let the user browse through these clusters to choose a starting point.

3.2.2. *The Manual Query System*

As baseline system, we used the *Manual Query System* (MQS) (Fig. 2(a)) resembling a 'traditional' image retrieval system, which returns a set of relevant images in response to a user-given query. A query can be formulated by a set of keywords, and/or one or more images as 'visual examples'. The user can also set the weighting between the two features. If not satisfied with the results returned by the system, the user has to alter their query and so forth.

4. Query Adaptation Techniques

In the course of a search session, a user creates and moves along a path of images. During each iteration, the path changes and the query needs to be adapted accordingly. The selected documents are treated as evidence of the user's information need, with a changing degree of uncertainty associated to each document: the older the evidence, the more uncertain we are that it is still indicative of the *current* information need. The degree of uncertainty is represented by an *ostensive relevance profile* [2], used to weigh the contribution of each path document. A separate query is constructed for each feature as a weighted combination of the documents' features.

Text Query: A new text query vector is created by updating the term's intrinsic weights (e.g. inverse document frequency (idf)) with the ostensive relevance weights resulting from the ostensive profile. The query vector then consists of the union of the set of terms that appear in any of the documents in the path. A term's original weight is multiplied by the sum of ostensive relevance values for all documents in which the term appears:

$$w_t = idf_t \times \sum_{\substack{i=1 \\ t \,\in\, D_i}}^{l_p} (ORel_i \;\times\; tf_t(D_i)) \tag{1}$$

where w_t is the resulting weight of term t in the query vector, idf_t the term's idf value, l_p the length of the path, D_i the document at position i in the path (starting at the most recently selected object), $tf_t(D_i)$ the term frequency of term t in document D_i, and $ORel_i$ the ostensive relevance weight at position i. The ostensive relevance weights are computed by the relevance profile function $\frac{1}{2^i}$, normalised to sum to 1: $\sum_{i=1}^{l_p} ORel_i = 1$.

Hence, the query terms are weighed with respect to the relevance profile and their corresponding *idf* values. A new query vector is computed based on the four highest ranking terms.

Histogram Query: There are different techniques for combining the query histogram from the individual histograms. A straight-forward approach in accordance with other query-point movement techniques (e.g. [18]) is a linear combination of the constituent histograms and the ostensive relevance weights:

$$H_Q = \sum_{i=1}^{l_p} (ORel_i \times H_{D_i}) \tag{2}$$

The resulting query histogram H_Q is comprised of the bins computed as the weighted sum of the path documents' bins. It can be interpreted as the weighted 'centroid' of the corresponding histograms.

4.1. Final Evidence

Two queries representing each feature are issued to the system, returning two result lists with different scores based on the respective similarity measure for each feature. For this reason, a means to combine the results to obtain one single ranked list of documents needs to be found. The *Dempster-Shafer Theory of Evidence Combination* provides a powerful framework for this combination.

The Dempster-Shafer mechanism has been widely used in the context of IR to combine information from multiple sources [14, 12]. The advantage of Dempster's combination rule, is that it integrates degrees of uncertainties or trust values for different sources. For two features Dempster-Shafer's formula is given by:

$$\begin{aligned} m(\{d_i\}) &= m_1(\{d_i\}) \times m_2(\{d_i\}) + \\ & \quad m_1(\Theta) \times m_2(\{d_i\}) + m_1(\{d_i\}) \times m_2(\Theta) \end{aligned} \tag{3}$$

and

$$m(\Theta) = m_1(\Theta) \times m_2(\Theta) \tag{4}$$

where $m_k(\{d_i\})$ (for $k = 1, 2$) can be interpreted as the probability that document d_i is relevant with respect to source k. The two sources in our case correspond to the similarity values computed from the text and colour feature respectively. Θ denotes the global set of documents, and $m_k(\Theta)$ represents the uncertainty in those sources of evidence (also referred to as un-trust coefficients):

$$m_k(\Theta) = 1 - strength_k \tag{5}$$

where:

$$strength_k = \frac{\sum_{i=1}^{l_p} m_k(\{d_i\})}{\sum_{i=1}^{l_p} m_1(\{d_i\}) + \sum_{i=1}^{l_p} m_2(\{d_i\})} \tag{6}$$

$strength_k$ corresponds to the trust in a source of evidence k. This reflects the contribution of a given source in selecting that particular image. In our definition, it reflects the importance of each feature.

The piece of evidence in this case is the calculated similarity values for the two features $m_1(\{d_i\})$ and $m_2(\{d_i\})$. The resulting $m(\{d_i\})$ is thus the combined belief for document d_i. Formulae 3 & 4 are a simplified version of Dempster-Shafer theory for IR purposes. Furthermore, it can easily be extended to accommodate more than two sources.

5. Experimental Methodology

Evaluation of interactive systems is a difficult problem. It has been argued that traditional IR evaluation techniques based on precision-recall measures are not suitable for evaluating adaptive systems [10, 13, 7, 16]. Two of the most important reasons are the subjectivity of relevance judgements on the one hand, and the importance of usability for a system's overall effectiveness, on the other hand. Usability can only be measured with the user in the loop, and will give valuable insights in what the users actually do rather than what we expect them to do. In addition, precision and recall can measure the effectiveness of the underlying algorithm relying on relevance judgments. However, relevance depends on a number of factors, such as topic, task, and context [7].

Further, it has been observed that image content is especially subjective and dependent on an individual's interpretation (e.g. [25, 23]).

The results of the evaluation reported in this paper show evidence on the subjectivity of relevance, even when task context is taken into consideration. For each of the three tasks we set, we asked the user to select all suitable candidate images before making a final choice while searching the collection (more information on the nature of tasks follows below in Section 5.1). If we determine the number of relevant images for a task by taking the union of selected candidates of all users, we obtain 76, 80, and 82 for the three tasks respectively. Counting the number of users that selected the candidates reveals that a huge majority has only been selected by one user, 67%, 71% and 65% respectively, showing that there is not very much user consensus. The percentages of candidates chosen by three or more users are 6.6%, 12.5% and 13.4%.

Hence, in order to evaluate our approach, we used a work-task oriented, user-centred approach, similar to that described in [13] with major emphasis on the usability of the interface. In our evaluative study, we adopted a randomised within-subjects design in which 18 searchers used three systems on three tasks. The independent variable was system type; three sets of values of a variety of dependent variables indicative of acceptability or user satisfaction were to be determined through the administration of questionnaires. To reduce the effect of learning from one system to the other, the order of the systems and tasks was rotated according to a Greco-Latin square design. The searches were performed on a collection containing 800 photographs, created from the photographic archive of the National Trust for Scotland [12].

5.1. Tasks

In order to place our participants in a real work task scenario, we used simulated work task situation [13]. This scenario allows users to evolve their information needs in just the same dynamic manner as such needs might develop during a 'real' retrieval session (as part of their normal work tasks). Before starting the evaluation, the users were presented with the work task scenario (see Fig. 3). For each system, they were given a different topic for the work task, each involving at least two searches. The topics were chosen to be of very similar nature, in order to minimise bias in the performance across the systems.

5.2. Systems

The Ostensive Browsers (Section 3.2.1) were evaluated against the 'traditional' image retrieval system MQS (Section 3.2.2), which supports manual query facilities. The Ostensive Browsers vary in the amount of

> *Imagine you are a designer with responsibility for the design of leaflets on various subjects for the Scottish Tourist Board [...]. These leaflets [...] consisting of a body of text interspersed with up to 4–5 images selected on the basis of appropriateness to the use to which the leaflets are put.*
>
> *Your task is to make a selection, from a large collection of images, of those that in your opinion would most effectively support the given topic. In order to perform this task, you have the opportunity to make use of a computerised image retrieval system, the operation of which will be demonstrated to you.*

Figure 3. Task Description

control options granted to the user. The *Pure Ostensive Browser* (POB) relies only on automatic query adaptation as described in Section 4, whereas the *Controlled Ostensive Browser* (COB) additionally provides options for selecting the features and their associated weights.

5.3. Hypothesis

Our experimental hypothesis is that the ostensive approach (reflected in both POB and COB) is generally more acceptable or satisfying to the user. It can be further distinguished in two sub-hypotheses: (1) Query adaptation coupled with an ostensive interface provides a better environment for CBIR, and (2) Providing an explicit control on the ostensive system results in better satisfaction on task completion.

5.4. Participants

In order to obtain data as close to real-life usage as possible, we sought design professionals as participants. Our sample user population consisted of 18 post-graduate design students. We met each participant separately and followed the procedure outlined below:

- an introductory orientation session
- a pre-search questionnaire
- for each of the three systems in turn:
  - a training session on the system
  - a hand-out of written instructions for the task
  - a search session in which the user interacted with the system in pursuit of the task
  - a post-search questionnaire
- a final questionnaire

We did not impose a time limit on the individual search sessions. The complete experiment took between 1.5h and 2h, depending on the time a participant spent on searching.

6. Results Analysis

6.1. Pre-search Questionnaire

Through this questionnaire, information about the participants' experience with computers and familiarity with using photographs was obtained. The participants were students at a post-graduate level in a design related field (graphic design, photography, architecture). Their ages ranged between 23 and 30 years. The ratio between male and female participants was (approximately) 2:1. The responses revealed that all of the users employed images extensively for their work, and that they are often required to retrieve images from large collections.

In summary, results from this questionnaire indicated that our participants could be assumed to have a good understanding of the design task we were to set them, but a more limited knowledge or experience of the search process. We could also safely assume that they had no prior knowledge of the experimental systems. The participants' responses thus confirmed that they were from the expected user population for the design task using an automated image retrieval system.

6.2. Post-search Questionnaire

After completing a search session on one of the systems given a particular task, the users were asked to complete a questionnaire about their search experience.

6.2.1. *Semantic Differentials*

Each respondent was asked to describe various aspects of their experience of using each system, by scoring each system on the same set of 28 7-point semantic differentials. The differentials focused on five different aspects (see Table 2):

- three of these differentials focused on the *task* set (Part 1);
- six focused on the *search process* that the respondent had just carried out (Part 2);
- five focused on the set of images *retrieved* (Part 3);
- three focused on the user's perception of the *interaction* with the system (Part 4); and
- eleven focused on the *system* itself (Part 5).

Table 2. Semantic differentials

Was the *task*...?
(clear↔unclear), (simple↔complex), (familiar↔unfamiliar)
Was the *search process*...?
(relaxing↔stressful), (interesting↔boring), (restful↔tiring), (easy↔difficult), (simple↔complex), (pleasant↔unpleasant)
Was the *retrieved set*...?
(relevant↔irrelevant), (important↔unimportant), (useful↔useless), (appropriate↔inappropriate), (complete↔incomplete)
Did you *feel*...?
(in control↔lost), (comfortable↔uncomfortable), (confident↔unconfident)
Was the *system*...?
(efficient↔inefficient), (satisfying↔frustrating), (reliable↔unreliable), (flexible↔rigid), (useful↔useless), (easy↔difficult), (novel↔standard), (fast↔slow), (simple↔complex), (stimulating↔ dull), (effective↔ineffective)

The result was a set of 1512 scores on a scale of 1 to 7: 18 respondents scoring each of 3 systems on 28 differentials. On the questionnaire form, the arrangement of positive and negative descriptors was randomised.

In our within-subject design, the sets of 18 scores on each differential for the three systems were compared. Our experimental hypothesis was that, in any individual case, the set of scores for both COB and POB was drawn from a population of lower (better) scores than that for MQS, and that COB scores were slightly lower than POB scores. Given the ordinal scale of the data, we had to use rank-based statistics. Since the data were not normally distributed, we calculated values of the non-parametric form of analysis of variance—the Friedman test. The null hypothesis in this case is: there is no difference in median ranks between groups on the criterion variable.

Overall, the Ostensive Browsers outperformed MQS, and usually COB's scores were lower (better) than the scores for its pure counterpart. The means of all differentials for each part is depicted in Fig. 4, which shows the trend that MQS scores are poorer than the scores for the other two systems, supporting our initial claim that query adaptation along with an ostensive interface provided a better environment for CBIR. The graph also shows quite clearly that POB's scores are comparable with COB, apart from the scores for Part 3 (images). This

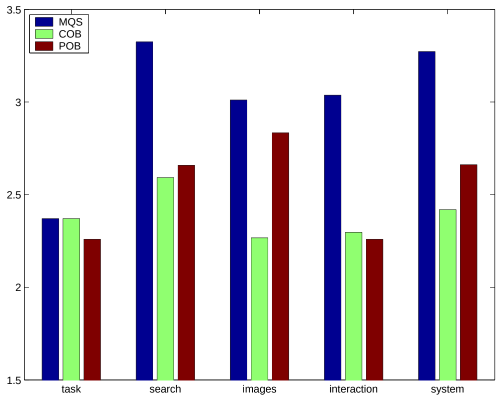


Figure 4. Semantic differential means per part (value range 1-7, lower=better)

part focused on the retrieved images, thus backing up our second subhypothesis, namely that providing an explicit control on the ostensive system resulted in better satisfaction on the task completion.

For each differential, we tested the hypothesis that the scores for each system type were sampled from different populations. The subset of differentials, which showed a significant level at $p <= 0.05$, are (p–value after adjustment for ties): 'restful' (0.008), 'pleasant' (0.05); 'comfortable' (0.014); 'flexible' (0.007), 'useful' (0.001), 'novel' (0.01), 'simple' (0.03), 'stimulating' (0.003) and 'effective' (0.007). Dunn's multiple comparison post test was performed to determine between which of the systems the difference occurred. For most differentials the significant difference occurred between MQS and COB. For 'pleasant', 'comfortable', and 'simple' no such matching could be found, however. The most significant results are found when comparing the differentials for the system part (Part 5). Most notable is the variance in judging the system's usefulness (the mean scores were 3.4, 2.6, and 1.9 for MQS, POB and COB, respectively), and it should be pointed out that the advantage of the POB as being the simplest tool to use is reflected in the results, as well (2.9, 2.2, and 2.9). A table showing these results in more detail can be found in [31].

There were no significant differences for Part 1 (concerning the tasks), neither across the systems nor across the tasks, which shows that the tasks were well-balanced and are believed not to have confounded the results significantly.

6.2.2. *Likert Scales*

Each user was asked to indicate, by making a selection from a 7-point Likert scale, the degree to which they agreed or disagreed with each of seven statements about various aspects of the search process and their interaction with the system. There were four statements concerning the *user's information need*. They were phrased in such a way that responses would indicate the extent to which:

1. the user's initial information need was well-defined ("I had an idea of the kind of images that would satisfy my requirement before starting the search.");
2. the user was able to find images representative or coextensive with that need ("The retrieved images match my initial idea very closely.")
3. the user's information need changed in the course of the search ("I frequently changed my mind on the images that I was looking for.");
4. the change of his need was due to the facilities offered by the system ("Working through the image browser gave me alternate ideas.").

The remaining statements captured the *user's satisfaction* with the search process and the system. Their responses would indicate the extent to which the user was satisfied with:

5. the outcome of their search ("I am very happy with the images I chose.");
6. the level of *recall* attained ("I believe that I have seen all the possible images that satisfy my requirement.");
7. the overall outcome of their interaction with the system ("I believe I have succeeded in my performance of the design task.").

Like before, each user was asked to respond to these statements three times (after each task they carried out on the different systems). The result was a set of 378 scores on a scale of 1 to 7 (with 1 representing the response "I agree completely" and 7 representing the response "I disagree completely"): 18 respondents scoring each of three systems with respect to each of the seven statements. The mean results are shown in the table in Figure 5(a).

Furthermore, since an evaluation based on the retrieved images *after* the search had been completed is hindered by subjective bias [1], the participants were invited to draw sketches of the kind of images they had in mind before starting the search (if they had any). This ensured that there was a point of reference for them to judge whether the retrieved images matched their initial sketches.

6.2.2.1. *Information Need Development:* The scores for the respondents' reactions to the statements regarding their information need requires careful consideration. When they were asked about their initial idea of the images they were looking for, the responses showed that

Stmt.	MQS	POB	COB
1	1.8	1.4	2.2
2	3.2	3.0	3.2
3	4.2	4.0	3.4
4	3.3	3.0	2.4
5	2.8	2.7	2.3
6	4.1	3.0	2.9
7	2.9	2.4	2.3

(a) Means for each statement

	MQS	COB
Statement 2		
images don't match initial idea	4	5
Statement 3		
changed mind on images	8	9
didn't change mind	7	4

(b) Split of answers on changing ideas. (Number of responses per statement.)

Figure 5. Tables for Likert Scale results

their initial need was reasonably well-defined (Stmt. 1). Users of COB were more inclined to change the initial need than for MQS and POB (Stmt. 3). However, the responses for the second statement whether the retrieved images matched their initial information need, were uniform across the systems (Stmt. 2). Still, when asked whether they thought the system gave them alternate ideas, COB scored significantly better (Stmt. 4). The significance of the difference is reflected in the values of the Friedman test statistics calculated in order to test the experimental hypothesis that the scores for COB are better (lower) than for MQS. The value of the Friedman statistic was found to be significant at a level of $p < 0.05$ ($p = 0.024$).

Analysing the comments about why they thought the images matched their initial idea (Stmt. 2) and why they changed their idea (Stmt. 3) sheds more light on the above results. We split the responses for these two statements into two categories: either their initial idea changed or did not change. For each category we considered only the responses where people stated they agreed (answers on the scale of 1–3) or disagreed (5–7). The table in Fig. 5(b) shows the resulting split of answers.

A comparison of the responses for MQS and COB yields the following results (responses for POB are very similar to those for COB and are therefore omitted). For MQS, all of the 4 users, who believed that the retrieved images do not match their initial idea (Stmt. 2), indicated that was because they could not find the images they visualised: *"I could not find the right ones"* or *"the system gives you slightly unexpected results"*. The same reasons were also brought forward by 4 out of 8 users when asked about their opinion of why they changed their mind (Stmt. 3).

On the contrary, the comments for COB suggest that 4 out of 5 people deviated from their initial idea rather because they were offered a bigger choice and variety in the images: *"there were plenty of images to choose from"* or *"I found other cool images first"*. 7 out of a total of 9 who changed their mind in COB thought this was the case because they were offered a better selection of images: *"the idea of having related images displayed next to each other evokes reconsideration of choices or sparks off other ideas. It makes it easier to choose between images"* also showing the advantages of the presentation of the retrieved images. These comments highlight the reasons for changes in their information need in the course of the interaction with the system. Therefore there is a necessity for system adaptation to reflect the changing needs.

A similar comparison can be made for the users' judgements of why they thought they did *not* change their minds. All 4 users who indicated that their information need remained constant on COB stated that they just had a clear idea of what kind of images they wanted: *"got more of the images I wanted"*. The reasons of why it remained constant on MQS are quite different. Only 2 out of 7 people in total who claimed they did not change their mind believed that they had a clear image: *"had ideas and stuck to them"*. 4 users however pointed the reason to the missing option of exploring the database: *"I saw less images–could not explore lines of images"* and *"more direct way of searching not leaving as many images to choose from"*.

To summarise, from the above analysis it emerges that, while most users had a mental model of candidate images, this model was changing during the search process. The system used had a major impact on the reasons for such changes. COB supported an explorative causing their needs to evolve by offering a large selection of alternative choices. In MQS however, many people at some point faced the problem that they were unable to retrieve any more images (usually when they exhausted keywords). They often had the feeling that the images they were looking for were not in the database, and they were often puzzled and frustrated because they could not tell whether the images were indeed not there or whether they could not formulate the query properly. The majority of people who changed their mind on the initial images interpreted that in a negative way as a result of not being able to find the right ones. One person's comment reflects this mood: *"My expectations have been adapted to the available images. This, however, is not how a designer wants to think, he doesn't want limitations to influence decisions."*

6.2.2.2. *User Satisfaction* When analysing the scores of the statements concerned with the overall user satisfaction, no significant differences could be shown to conclude an overall improvement on satisfaction on task completion. Still, MQS's scores were always poorer, and the user comments presented below support the observation that they were generally more happy with the selection of images in the browsers. There are various other factors that can influence the satisfaction on task completion, too, for example the available images in the collection. After all, if a user is not really happy with the available images, none of the three system would be able to change this. Due to the relatively small sample size in our study, only a small number of such outliers can have an effect on the statistical significance of the results.

6.2.3. *Open Questions*

In order to gain more insight into the users' preferences, the participants were asked to specify which features of the system they liked, which ones they disliked, and what features they would like to have seen added. The responses obtained here were quite similar to the ones in the final questionnaire. To avoid repetition, they are presented together in the next section.

6.3. Final Questionnaire

After having completed all three tasks, the participants were asked to rank the three systems in order of preference with respect to (i) the one that *helped* more in the execution of their task, and (ii) the one they *liked* best. Both questions resulted in a very similar ranking of the systems. 10 out of the 18 participants ranked COB more highly than the other systems, and 12 placed both ostensive interfaces as their top two. The mean of the ranks were: MQS 2.5, POB 1.9, and COB 1.6. Again, in order to test the experimental hypothesis that the sets of 18 post-search scores for each system type were sampled from different populations, the Friedman statistic was calculated, which was found to be significant at a level of $p = 0.017$ (for both Questions (i) and (ii)). Dunn's post test showed that a significant difference was between MQS and COB (with $p <= 0.05$), however not between MQS and POB. Our conclusion, therefore, was that people liked COB significantly better than MQS, and found it significantly more useful for the task we set them.

Respondents who ranked MQS highest appreciated the system's accuracy and being able to control the search—e.g. *"fastest of the 3 systems in finding specific images"*. The features liked most were the combination of visual and textual features. However, some users found it hard to interpret the features and how to specify the correct combination. The complexity of formulating a query in MQS emerged in many comments: *"quite complex"*, *"have to input too often"*, *"confus-*

ing to control". Some people also found MQS *"too restrictive"*. Other participants, who used one of the other systems first, missed the ability to browse the images or return to previously retrieved images.

Those respondents who preferred either of the ostensive browsing approaches valued the fact that they were very intuitive and appreciated the *"visual representation of the search process"* (*"easily understandable 'line of choices'"*, *"ability to compare images on screen + backtracking"*). They considered the search process a *"very image-based selection"*. The difference between the two approaches seems to be the flexibility on the one hand (COB) and the ease of use (POB) on the other hand. POB was generally referred to as *"very intuitive, fast"*, *"pleasure to use"* and *"relaxing"*. Arguments for the POB approach included: *"it is easier to pick images rather than to choose words"* and *"very fluid movement—just the images"*. POB's drawbacks were concentrated on the missing ability to control the search (*"being stuck in a sequence, not being able to edit and control it"*). The additional control options, however, were also criticised by some users in COB. Few people disliked the system's automatic selection or found it *"offered too much control, there's too much to think about"*.

Apart from this, most responses about COB were entirely positive. It was still deemed *"easy to understand"* and *"very straight-forward"*. In addition, people liked its adaptability and versatility. They seemed to consider this system a more complete approach (*"most options, best display of information"*) and regarded the system *"very helpful"* and *"intelligent"* in making *"smart selections"*. The effectiveness of the system is reflected in a lot of responses: *"it is most efficient to use and get the desired results"*, *"search seemed more consistent"*, and *"felt more extensive"*. Hardly anyone ever got stuck during the search process, and one of the features liked best included *"the fact that it kept going and going"*.

6.4. Quantitative Results

In order to test the actual user performance in quantifiable, objective measures, a number of usage data was logged during the experiments. The kind of data logged included:

- time taken for the complete search session
- number of distinct images retrieved during the search session
- number of searches per session

Most interestingly, the time taken for the whole session was not significantly different between the systems. On average the completion times were 15m20s for MQS, 15m30s for COB, and 13m54s for POB.

Comparing the individual times for each user it emerged that the completion time largely depended on the user: people tended to spend approximately the same amount of time for each system. A further factor is tiredness or boredom that might have affected the timing. The ordering of the systems had a slight impact on the time spent for searching: 16m29s for the first system used, 14min43sec for the second, and 14m31s for the third. Again, the differences are not large enough in order to conclude that tiredness influenced the evaluation adversely.

In contrast, the number of distinct images retrieved in approximately the same time span was much higher in the browsing systems. On average the number of distinct images for MQS, COB, and POB were 58.2, 82.9, and 83.0 respectively. The difference between MQS and COB could shown to be statistically significant ($p = 0.012$, value of Friedman statistic, adjusted for ties).

This is an indication that the browsing systems, by relieving the user from having to formulate explicit queries, succeeds in the user seeing more images in the collection. We believe that the time the user has to spend on the query formulation and re-formulation in MQS is used in a more productive way in the browsers. In fact, POB (in which there is no query formulation process necessary at all) has the highest rate of image recall per minute (6.0 compared to 5.4 for COB and 3.8 for MQS).

6.5. Observations

The observations of the participants using the system revealed further interesting facts. One—probably the most prevalent—issue to arise was the problems associated with the use of keywords. First of all, only few people used more than one search term at a time. Furthermore, they were often surprised about the results they obtained. The subjectivity of the choice of terms to describe an image was apparent throughout (*"summer? that's not my definition of summer!"*). This was especially limiting in the MQS, since the keyword search was the mostly used feature in this system. As a result, most people considered the task to be finished after they could not think of any more keywords to use. The only option of retrieving more images in the MQS, when the user exhausted words, was to play around with the weighting between the two features. Many participants took this approach, but it was more a trial-and-error process in order to see whether they can retrieve any different images.

Another problem, which became most apparent in the MQS was that people cannot easily relate to content-based image features. Even though they were told that the feature used was 'colour only', most people when selecting ‘query-by-example’ representations, had the idea set in their mind that they wanted ‘more images like this one’. They could not distinguish between ‘images that have the same colour’ and ‘images that are generally similar’ (in terms of semantic content, layout, colour, etc.). As a result, they often obtained unexpected results, since the returned images did not resemble—in their minds—the ‘query-by-example’ images.

A further interesting point to notice is that the ostensive browsing approaches seemed to be more successful in giving the users confidence and insight in the available images. The users got the perception that *“there are so many more images to choose from”*. On the other hand, when using the MQS people thought the image collection to be *“limiting”*. In addition, people often felt lost, because they could not tell whether they were not able to formulate the query or whether the images were simply not in the collection. While using the browsing systems, this uncertainty did not arise due to the different approach to searching. The perception of the participants is reflected in numbers: the number of distinct images seen was indeed higher for the browsing approaches (see Section 6.4).

Most of those points mentioned are subjective observations of the evaluators only, and cannot be shown to be representative or ‘statistically significant’. However, they convey interesting aspects and are believed to help in the further design of image retrieval systems. Moreover, they are in accordance with other studies [13, 15, 16].

7. Discussion

First of all, many problems of using a traditional CBIR system became evident in the evaluation presented above. These included the difficulty posed for the user to interpret low-level features and its effects on the query formulation, and the uncertainty the users felt about the availability of images and their ability to retrieve them. Our approach supports a way of adaptive *content-assisted browsing*, addressing many of these difficulties the user has to face in an image retrieval system. The user is not required to explicitly formulate their need as a query, which is instead incrementally constructed by the system based on the user’s choice of images.

The evaluation showed that people preferred the search process in the ostensive browsing scheme, felt more comfortable during the interaction, and generally found the system more satisfactory to use compared to the traditional CBIR interface (see Section 6.2.1). In a study concerning the nature of the information need, it emerged that the Ostensive Browser (OB) provides for an explorative search that reflects and supports dynamically changing needs (see Section 6.2.2). The analysis of user comments supported the view that the user's underlying need changes while they explore the collection, although they mostly have a mental model before starting the search. Our approach was more successful in eliciting such changes and adapting the retrieval along these lines. This defends our proposition for an adaptive, interactive retrieval system. The two versions of the OB we provided revealed a tradeoff between simplicity (for the pure version) and flexibility (by providing additional control facilities). While most participants preferred the flexibility, they also appreciated the pure browser's simplicity.

Since there are many different types of users and different types of searches (well-defined, ill-defined, searching for a known object, searching to get inspiration, etc.), the retrieval system should attempt to cater for these variations. We believe the evaluation proved the success towards a consistent, effective and versatile approach in the form of the OB equipped with additional control facilities. It provides a simple browsing-like interaction that allows for an explorative search and serendipitous discovery. The adaptive scheme emulates the development of the user's need during such explorative phases. When the user is more clear about the search, the additional control facilities permit the user to take over and steer the retrieval in a certain direction. Again, a user's comment summarises this ability: *"I liked the flexibility when I needed and the automatic selection when I didn't"*.

Evaluation in image retrieval systems is a difficult task. Traditional techniques based on precision-recall measures evaluating the retrieval effectiveness have often been criticised for treating the system as an independent entity [10, 13, 16]. However, it has been recognised that the user plays a vital role in the design and evaluation of CBIR systems. Since image retrieval is an inherently interactive activity, a user-centred evaluation, in which 'real' people use the system in a 'real-world' setting, can provide invaluable insights in the system's overall performance that precision-recall measures can never anticipate. Important performance indicators ignored in traditional evaluations are user interface issues, task completion time and user satisfaction.

For this reason, we designed our evaluation to follow the guidelines of the evaluative framework for interactive, multimedia retrieval systems proposed in [13]. The main points in our evaluation following these guidelines are:

- Design professionals were asked to participate in the study in order to test the systems with real potential users.

- Context-situated tasks were created to place the participants in a 'real-life' usage scenario.
- A variety of qualitative measures indicative of user satisfaction (concerning the system, the tasks, the interface, etc.) was collected and analysed.
- Quantitative measures on task-completion time and images retrieved confirmed the qualitative measures of user satisfaction.

The image collection used is relatively small and might have had an impact on practice effects. Although the choice of topics for the tasks was such that the overlap of images suitable for each topic was minimised, further studies are needed with a much larger collection in order to generalise the results. In the meantime, however, the results discussed above highlight many important aspects of CBIR interface design. While the usability of a system depends largely on its interface, the performance of the underlying algorithms cannot be neglected for judging a system's overall effectiveness. The retrieval performance of the OM-based query learning scheme is better judged in comparison to other relevance feedback techniques in a more objective quantitative evaluation. In the following, we will discuss results from a preliminary quantitative evaluation to this end.

8. Preliminary Study on Retrieval Performance of OM

We have set up a simulated comparative evaluation to measure the retrieval performance of the OM. In this experiment, we are interested in how well the OM performs in terms of the number of images found in a category search. The number of relevant images retrieved is an indication of the overall level of recall, i.e. the number of relevant images retrieved divided by the total number of relevant images for a category. The number of iterations until a session converges (when the system is not able to return any new relevant images) gives an indication of user effort to retrieve all these images. In the ideal case, while maximising recall the iteration number should be low, meaning that the retrieval system succeeds in returning all relevant images early in the session.

The query learning scheme proposed by Rui et al. [20] serves as baseline. Rui's scheme is essentially a relevance feedback technique, which represents a query as the average over all positive examples in addition to a feature re-weighting scheme. To make the comparison fair, the same query learning and feature re-weighting is employed for the OM. The idea behind the learning scheme is still the same as the one proposed in Section 4. Similar to the query representation in Equation 2, the new query is computed as the *weighted* (with the ostensive relevance weights) average of the path images in the OM. Instead of using the Dempster-Shafer theory, however, the features are linearly combined using the feature weights computed according to Rui's scheme. The details of the feature representation and re-weighting scheme can be found in [20].

The evaluation is performed on a subset of the Corel dataset (Photo CD 4), containing 24 categories of 100 images each. Since we have no keyword data associated with this collection, we only use content-based features for this evaluation. The 6 low-level colour, texture and shape features implemented are (feature dimension): Average RGB (3), Colour Moments (9) [26]; Co-occurrence (20), Autocorrelation (25), and Edge Frequency (25) [24]; Invariant Moments (7) [9].

8.1. The Simulation Setup

We simulated user interaction to find as many relevant images from a given category as possible. An image is considered relevant if it belongs to the same category as the initial query image. The simulation for the baseline system is as follows. Starting with one image from the given category, the system returns the 20 most similar images (images already in the query are not returned again so as to maximise the system's ability of collecting a large number of distinct relevant images). From this set, the simulated user selects at most n relevant images to add to the query and the system recomputes the top 20 images. The process is iterated until there are no more relevant images in the returned set. This simulation resembles the traditional relevance feedback process. We report results from two variations of n: a "realistic" scenario where n=3, referred to as RFS$_3$, and a "greedy" scenario, RFS$_g$, where n=20.

The simulation setup for the OM is slightly different. Starting with one query image as the root image, the system returns the top k ($6 \leq k \leq 12$) candidates. The user selects the first relevant image from the returned set, and adds it to the path. The process is repeated until there are no more relevant images in the latest candidates. At this point, the user backs up along the path and continues with the closest image, which has unprocessed relevant candidates. This corresponds to a depth-first traversal of the ostensive tree. The session continues until there are no more new relevant images in the ostensive tree. There are two assumptions being made about the user's actions. First, the simulated user only selects relevant images and second, once a relevant image has been selected in one path, it will not be pursued in a different branch again. (There can be duplicate images in different branches of the ostensive tree, even though the candidates will not contain any image already on the current path.) The simulation scheme will be referred to as OMS_k.

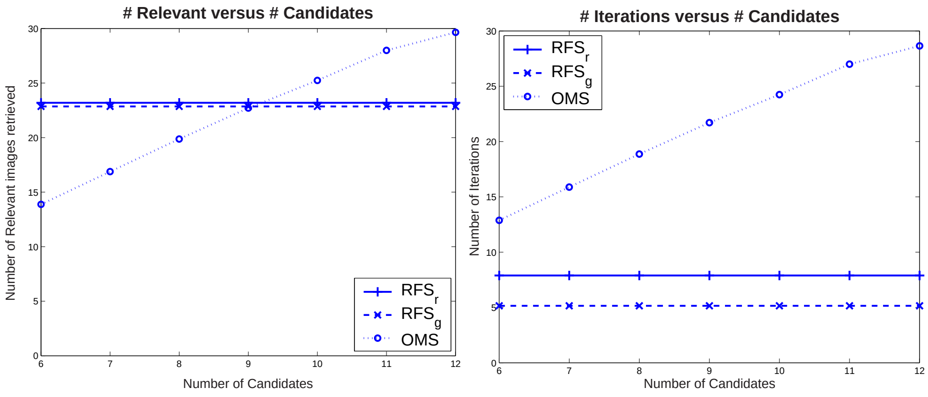
8.2. Results Analysis

We have chosen five categories that contain visually similar images. Other categories are very difficult for CBIR, so that the majority of queries would not return any relevant images in the first iteration. The selected categories are: 'lions', 'elephants', 'horses', 'mountains', 'orchids'. Every image in each category serves as the first query image for both schemes, RFS and OMS, resulting in a total of 100 queries per category.

Figure 6(a) shows the average number of (unique) relevant images found for all five categories in RFS_3, RFS_g as well as OMS_k for various candidate sizes k. The performance of RFS_3 and RFS_g is very similar, with RFS_3 results being slightly better. It can be seen that for $k = 9$, OMS_9 succeeds in finding approximately the same number of relevant images as both RF scenarios. Increasing k results in OMS_k outperforming RFS in terms of the level of recall. (An example of how to display a larger number of candidates in the interface is displayed in Figure 7.) However, as the number of candidates increases and more relevant images can be found, the number of iterations until convergence increases with it. (Note, that in the OM simulation the iteration number is always one more than the number of unique relevant images found, since each relevant image will be selected exactly once, plus one for the final iteration, which fails to return any relevant images). The iteration number of OMS_6 is already higher than the baseline, as can be seen in Figure 6(b). Table 3 summarises these results. It can be seen that RFS_3 converges after approx. 8 iterations, while the greedy strategy only needs 5 iterations to achieve a similar level of recall. Although RFS apparently converges faster than OMS, this does not necessarily mean that RFS requires less user effort (in terms of mouse clicks for example). Keeping in mind that the number of relevant images for feedback in RFS is n, i.e. the user has to click up to n+1 times (n for feedback, 1 for new search) whereas in the OM scenario only 1 click is required to initiate a new iteration. The average click count for RFS_n in the simulation was 28, which interestingly is very similar to the iteration number of OMS_{12}.

8.3. Limitations of the Study

The evaluation presented is merely a preliminary study, and a lot of additional factors besides the recommendation size can be considered, such as for example the choice of ostensive relevance profile (see



(a) Nr. of Relevant images (b) Nr. of Iterations

Figure 6. Nr. of Relevant images and nr. of Iterations vs candidate size.

Table 3. Average results for nr. of relevant images retrieved (R), nr. of iterations (I) and nr. of relevant per iteration (R/I).

	RFS$_3$	RFS$_g$	OMS$_6$	OM$_7$	OM$_8$	OM$_9$	OM$_{10}$	OM$_{11}$	OM$_{12}$
R	23.19	22.86	13.88	16.88	19.87	22.71	25.24	27.99	29.65
I	7.89	5.15	12.88	15.88	18.87	21.71	24.24	26.99	28.65
R/I	2.94	4.44	1.08	1.06	1.05	1.05	1.04	1.04	1.04

Section 4). Also our assumptions about the user's actions might not necessarily be realistic. Does a user always select relevant information? Does a user rather proceed deep along a path (depth-first traversal as modelled here) or rather select all relevant options first (breadth-first)? In a future study, a proper user model should be constructed or even better, real users should conduct the searches.

Another favourable point is that the interaction possibilities in an Ostensive Browser allow for more flexibility than in a traditional rele-

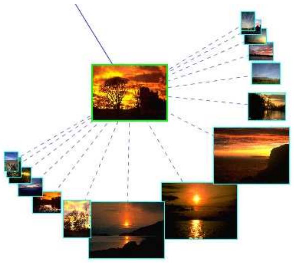


Figure 7. Example of fisheye display for candidate size of 15.

vance feedback system. In an RF scenario, all images selected relevant will be accumulated and added to the query. In contrast, the Ostensive Browser gives the user more control over which images are relevant for a query by moving back and forth in one path and branching off into different directions. The effect of the selection strategy is a very interesting point to consider. In a future evaluation we could compare various user models in the simulation.

9. Conclusions and Future Work

We developed and described an adaptive retrieval approach towards CBIR based on an innovative browsing scheme. This approach is based on the concept of Ostension. The underlying idea is to mine and interpret the information from the user's interaction in order to understand the user's needs. The system's interpretation is used for suggesting new images to the user. Both text and colour features were employed and combined using the Dempster-Shafer theory of evidence combination. A user-centred, work-task oriented evaluation demonstrated the value of our technique by comparing it to a traditional CBIR interface. It also showed how the development of the information need during the information seeking process affects and is affected by the search system. In addition, the OM-based query learning strategy showed favourable retrieval performance in comparison to a standard relevance feedback technique in a simulated quantitative evaluation.

Future work includes an integration of the ostensive approach into more sophisticated search and browsing strategies. We believe that it is essential to treat the search process as only part of the whole work process. One step towards this is to eliminate the distinction between *search* and *organisation*. More can be learnt from the user when combining the search and organisation process and it should lead to easier and more natural interaction with the system. Furthermore, the results of this initial study have to be verified in a larger scale evaluation involving long-term usage of the system for the real day-to-day tasks of the users. This is hoped to be realised with an integrated system involving some sort of ostension to exploit the advantages of an Ostensive Browser shown in this work.

References

[1] J. A. Black, Jr., G. Fahmy and S. Panchanathan, "A Method for Evaluating the Performance of Content-Based Image Retrieval Systems Based on

Subjectively Determined Similarity between Images," in Proc. Int. Conf. on Image and Video Retrieval, LNCS 2383, 2002, pp. 356–366.

[2] I. Campbell, "Interactive Evaluation of the Ostensive Model, using a new Test-collection of Images with Multiple Relevance Assessments," Journal of Information Retrieval Vol. 2(1), pp. 89–114, 2000a.

[3] I. Campbell, "The Ostensive Model of Developing Information Needs," Ph.D. thesis, University of Glasgow, 2000b.

[4] I. Campbell and C. J. van Rijsbergen, "The Ostensive Model of Developing Information Needs," in Proc. Int. Conf. on Conceptions of Library and Information Science, 1996, pp. 251–268.

[5] M. Chalmers, K. Rodden and D. Brodbeck, "The Order of Things: Activity-Centred Information Access," Computer Networks and ISDN Systems, Vol. 30(1–7), pp. 359–367, 1998.

[6] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas and P. N. Yianilos, "The Bayesian Image Retrieval Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments," IEEE Trans. Image Processing, Vol. 9, pp. 20–37, 2000.

[7] M. Dunlop, "Reflections on Mira: Interactive Evaluation in Information Retrieval," Journal of the American Society for Information Science, Vol. 51(14), 1269–1274, 2000.

[8] S. R. Garber and M. B. Grunes: 1992, "The Art of Search: A Study of Art Directors," in Proc. ACM Int. Conf. on Human Factors in Computing Systems (CHI'92), 1992, pp. 157–163.

[9] M.-K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Transactions on Information Theory, Vol. IT-8, 179–187, 1962.

[10] P. Ingwersen, Information Retrieval Interaction, Taylor Graham: London, 1992.

[11] Y. Ishikawa, R. Subramanya and C. Faloutsos: "MindReader: Querying Databases through Multiple Examples," in Proc. 24th Int. Conf. on Very Large Data Bases, 1998, pp. 218–227.

[12] J. M. Jose, "An Integrated Approach for Multimedia Information Retrieval," Ph.D. thesis, The Robert Gordon University, Aberdeen, 1998.

[13] J. M. Jose, J. Furner and D. J. Harper, "Spatial Querying for Image Retrieval: A User-oriented Evaluation," in: Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1998 pp. 232–240.

[14] J. M. Jose and D. J. Harper: 1997, "A Retrieval Mechanism for Semi-Structured Photographic Collections," in Proc. of the Int. Conf. on Database and Expert Systems Applications, 1997, pp. 276–292.

[15] M. Markkula and E. Sormunen, "End-User Searching Challenges Indexing Practices in the Digital Newspaper Photo Archive," Information Retrieval, Vol. 1(4), pp. 259–285, 2000.

[16] S. McDonald, T.-S. Lai and J. Tait, "Evaluating a Content Based Image Retrieval System," in Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001, pp. 232–240.

[17] J. Peng, B. Bhanu and S. Qing: 1999, "Probabilistic Feature Relevance Learning for Content-Based Image Retrieval," Computer Vision and Image Understanding, Vol. 75(1/2), pp. 150–164, 1999.

[18] K. Porkaew, K. Chakrabarti and S. Mehrotra, "Query Refinement for Multimedia Similarity Retrieval in MARS," in Proc. ACM Int. Conf. on Multimedia, 1999, pp. 235–238.

[19] J. J. Rocchio, “Relevance feedback in information retrieval,” in G. Salton (ed.), The SMART retrieval system: experiments in automatic document processing, Prentice-Hall: Englewood Cliffs, NJ, 1971, pp. 313–323.

[20] Y. Rui and T. S. Huang, “Optimizing Learning in Image Retrieval,” in IEEE Proc. Conf. on Computer Vision and Pattern Recognition, 2000, pp. 236–245.

[21] G. Salton and C. Buckley, “Improving retrieval performance by relevance feedback,” Journal of the American Society for Information Science, Vol. 41(4), pp. 288–297, 1990.

[22] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill: Tokio, 1983.

[23] S. Santini, A. Gupta, and R. Jain, “Emergent semantics through interaction in image databases,” IEEE Trans. Knowledge and Data Engineering, Vol. 13(3), pp. 337–351, 2001

[24] M. Sonka, V. Hlavac and R. Boyle, Image Processing, Analysis, and Machine Vision, Brooks and Cole Publishing, 2nd edition, 1998.

[25] D. M. Squire and T. Pun, “Assessing Agreement Between Human and Machine Clustering of Image Databases,” Pattern Recognition, Vol. 31(12), pp. 1905–1919, 1998.

[26] M. Stricker and M. Orengo, “Similarity of color images,” in Proc. SPIE: Storage and Retrieval for Image and Video Databases, 1995, pp. 381–392.

[27] M. J. Swain and D. H. Ballard, “Color indexing,” Int. Journal of Computer Vision (Kluwer Academic Publishers), Vol. 7(1), pp. 11–32, 1991.

[28] A. H. M. ter Hofstede, H. A. Proper and T. P. van der Weide, “Query formulation as an information retrieval problem’. The Computer Journal, Vol. 39(4), pp. 255–274, 1996.

[29] K. Tieu and P. Viola, “Boosting Image Retrieval,” in IEEE Proc. Conf. on Computer Vision and Pattern Recognition, 2000, pp. 228–235.

[30] S. Tong and E. Chang, “Support Vector Machine Active Learning for Image Retrieval,” in Proc. ACM Int. Conf. on Multimedia, 2001, pp. 107–118.

[31] J. Urban, J. M. Jose and C. J. van Rijsbergen, “An Adaptive Approach Towards Content-Based Image Retrieval,” in Proc. Int. Workshop on Content-Based Multimedia Indexing (CBMI’03), 2003, pp. 119–126.

[32] R. W. White, J. M. Jose and I. Ruthven, “An Approach for Implicitly Detecting Information Needs,” in Proc. Int. Conf. on Information and Knowledge Management, 2003, pp. 504-507.

[33] R. W. White, I. Ruthven and J. M. Jose, “Finding Relevant Documents using Top Ranking Sentences : An Evaluation of Two Alternative Schemes,” in Proc. Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2002, pp. 446–446.

[34] M. E. J. Wood, B. T. Thomas, and N. W. Campbell, “Iterative Refinement by Relevance Feedback in Content-Based Digital Image Retrieval,” in Proc. ACM Int. Conf. on Multimedia, 1998, pp. 13–20.

[35] X. S. Zhou and T. Huang, “Relevance Feedback in Image Retrieval: A Comprehensive Review,” ACM Multimedia Systems Journal, Vol. 8(6), pp. 536–544, 2003.