An Explorative Study of Interface Support for Image Searching

Jana Urban and Joemon M. Jose

Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK
{jana, jj}@dcs.gla.ac.uk

Abstract. In this paper we study interfaces for image retrieval systems. Current image retrieval interfaces are limited to providing query facilities and result presentation. The user can inspect the results and possibly provide feedback on their relevance for the current query. Our approach, in contrast, encourages the user to group and organise their search results and thus provide more fine-grained feedback for the system. It combines the search and management process, which – according to our hypothesis – helps the user to conceptualise their search tasks and to overcome the query formulation problem. An evaluation, involving young design-professionals and different types of information seeking scenarios, shows that the proposed approach succeeds in encouraging the user to conceptualise their tasks and that it leads to increased user satisfaction. However, it could not be shown to increase performance. We identify the problems in the current setup, which when eliminated should lead to more effective searching overall.

1 Introduction and Motivation

Content-based image retrieval (CBIR) systems have still not managed to find favour with the public even after more than a decade of research effort in the field. There are two main reasons for their lack of acceptability: first, the low-level features used to represent images in the system do not reflect the high-level concepts the user has in mind when looking at an image (*semantic gap*); and – partially due to this – the user tends to have major difficulties in formulating and communicating their information need effectively (*query formulation problem*).

The semantic gap is inherent to CBIR [1] and finding better feature representation has been at the core of CBIR research since the early stages. A large variety of features has been proposed over the course of time: from the initial and still widely used low-level features, such as colour and texture, e.g. [2], to more high-level techniques, such as visual templates [3], and finally a combination of visual cues and textual annotations to arrive at "semantic" features, e.g. [4]. Despite this, the current techniques have not succeeded in bridging the semantic gap. A contributing factor is the fact that an image's meaning is very subjective and context-dependent, which makes it difficult to find generic solutions that do not incorporate the users' opinions.

M. Detyniecki et al. (Eds.): AMR 2005, LNCS 3877, pp. 207–221, 2006.

The query formulation problem, on the other hand, has emerged as an IR problem in general [5]. The internal representation of documents is optimised for indexing efficiency and retrieval performance, but is, more often than not, rather alien to the user. The semantic gap only amplifies the problems associated with creating a meaningful query that fulfills a user's request.

Hence, improving the way images are represented is only part of the story. In order to assist the user in communicating their requests effectively, better interfaces are needed. The interface should provide a natural means to communicate information needs, should elicit and detect changes in a user's need while interacting with the system, and should in general engage the user in the task they want to solve rather than in the details of how the retrieval system works.

With these requirements in mind, we have proposed a system, EGO, that combines the search and the management process [6]. While searching for images, the creation of groupings of related images is supported, encouraging the user to break the task up into related facets to organise their ideas and concepts. The system can then assist the user by recommending relevant images for selected groups. This way, the user can concentrate on solving specific tasks rather than having to think about how to create a good query in accordance with the retrieval mechanism. It allows the user to interact more directly with the results in a way that is closer to their mental model of solving a search task.

In this paper we present an explorative study comparing two interfaces with respect to the support they offer the user to search for images and organise their results. Our aim is to collect evidence on whether the proposed system helps the user to conceptualise their search tasks. Further, we test our hypothesis that EGO helps to overcome the query formulation problem, since – relying on the in-built recommendation system – there is no need to create a query in order to initiate a search. We measure EGO's success in these two issues compared to a traditional relevance feedback system as a baseline. In the relevance feedback system, the user is given the option of selecting relevant images from the search results in order to improve the results in the next iteration. The evaluation is based on real users, performing practical and relevant tasks and captures a large amount of interaction data, which can be used in follow-up evaluations requiring a long-time involvement of the user.

The remainder of the paper is organised as follows. The interfaces used in the evaluation are described in Section 2. Section 3 sets out the experimental methodology, followed by a detailed analysis of the results and a summarising discussion in Sections 4 and 5. Finally, Section 6 concludes the paper.

2 The Interfaces

As a result of the above requirements, we have designed the EGO system. EGO is a personalised image management and retrieval tool that learns from and adapts to a user by the way they interact with the image collection. The high-level concepts of the EGO system are described in the context of other CBIR systems in [6]. In the experiment we evaluate a simplified version of its interface. A traditional relevance feedback interface serves as baseline.

2.1 Retrieval System

The underlying retrieval system is the same in both interfaces. It involves choosing an ideal query and learning the parameters of the matching function by the user provided examples.

Image Representation. The images are represented according to the hierarchical object model proposed in [7]. In this model an image is represented by a set of feature vectors, one for each distinct feature implemented, rather than a single stacked feature vector.

Implemented Features. We use the following 6 low-level colour, texture and shape features (feature dimension): Average RGB (3), Colour Moments (9) [8]; Co-occurrence (20), Autocorrelation (25), Edge Frequency (25) [9]; Invariant Moments (7) [10].

Distance Measure. The distance between an object x in the database and a given query representation q is computed in two steps. First, the individual feature distances g_i (for i in $1..I$, where I is the number of features) are computed by the generalised Euclidean distance,

$$g_i = (\boldsymbol{q_i} - \boldsymbol{x_i})^T W_i (\boldsymbol{q_i} - \boldsymbol{x_i}) \tag{1}$$

where $\boldsymbol{q_i}$ and $\boldsymbol{x_i}$ are the i-th feature vectors of the query q and the database object x respectively, and W_i the *feature transformation matrix* used for weighting the feature components. W_i is a $K_i \times K_i$ real symmetric full matrix, where K_i is the i-th feature dimension. The second step is then to combine the individual distances to arrive at a single distance value d. This is achieved by a linear combination between $\boldsymbol{g} = [g_1, ..., g_I]^T$ and a feature weight vector $\boldsymbol{u}$,

$$d = \boldsymbol{u}^T \boldsymbol{g} \tag{2}$$

The Recommendation System is based on a relevance feedback algorithm, that attempts to learn the best query representation and feature weighting for a selected group of images (positive training samples).

Learning the Feature Weights. We adopt the optimised framework for learning the feature weights proposed in [11]. Due to the hierarchical object model, it distinguishes between intra- and inter-feature weights. The optimal intra-feature component weights are given by an optimal feature space transformation matrix W_i. W_i is calculated as,

$$W_i = \det(C_i)^{\frac{1}{K_i}} C_i^{-1} \tag{3}$$

where C_i is the *weighted covariance matrix* of the N positive examples according to the i-th feature. W_i takes the form of a full matrix, if N is larger than the dimensionality of the feature, otherwise only the diagonal entries are considered. The optimal inter-feature weights $\boldsymbol{u} = [u_1, .., u_I]$ are the weights that best capture the inter-similarity between the training samples. The $\boldsymbol{u_i}$'s are solved by,

$$u_i = \sum_{j=1}^{I} \sqrt{\frac{f_j}{f_i}} \tag{4}$$

where $f_i = \sum_{n=1}^{N} g_{ni}$. The optimal intra-feature weights W_i and the optimal inter-feature weights u are used in Equations (1) and (2) respectively to calculate the total distance between a database object and the query representation.

Computing the Query Representation and Ranked Results. Our proposed learning scheme relies on a form of query expansion. The chosen query representation for a group is a multi-point query [12], whereby each query point represents one cluster of visually similar images in the group. The query points are selected as the image closest to each cluster centroid, and are weighted relative to the cluster size. When issuing the multi-point query to the system, a separate result list will be returned for each query point, which need to be combined. An investigation of several combination strategies [13] has led us to choosing a rank-based *voting approach (VA)*. Please refer to [13] and [6] for more details.

For the purpose of the evaluation however, we are simply computing one overall query representation as in [11]. This is mainly due to computational complexity (so as not to stretch the users' patience), but also due to some anomalies we found during the evaluation due to the clustering algorithm used.

2.2 Workspace Interface - WS

The combination of retrieval and management system is achieved by providing a workspace in the interface which allows the user to organise their search results. Images can be dragged onto the workspace from any of the other panels (or imported from outside the system) and organised into groups. The grouping of images can be achieved in an interactive fashion with the help of a recommendation system. For a selected group, the system can recommend new images based on their similarity with the images already in the group. The user then has the option of accepting any of the recommended images by dragging them into an existing group.

The interface used in the evaluation is a simplified version of that of the EGO system. EGO has some additional features for personalisation and can, in principle, accommodate any sort of query facility. Since our main objective in these experiments is to evaluate the usefulness of the workspace (and also to avoid biasing the participants by the naming of the experimental systems), this interface is referred to as the Workspace Interface (*WS*). The WS interface depicted in Figure 1 comprises the following components:

1. Given Items Panel: This panel contains a selection of images (three per task) provided for illustration purposes and can be used to bootstrap the search;
2. QBE Panel. This provides a basic query facility to search the database by allowing the user to compose a search request by adding example images to this panel. Clicking on the "Search" button in this panel will issue a search, which causes the system to automatically construct a query from the examples provided and compute the most similar images in the database.
3. Results Panel: The search results from a query constructed in the QBE panel will be displayed in this panel. Any of the returned images can be dragged onto the workspace to start organising the collection or into the QBE panel to change the current query.

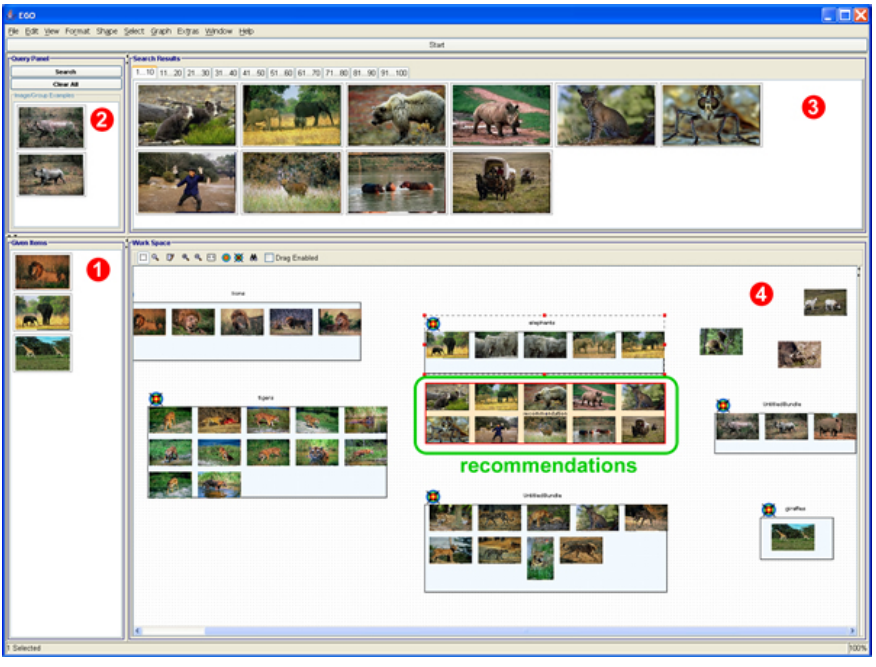

Fig. 1. Annotated WS interface

4. Workspace Panel: The workspace holds all the images added to it by the user, and serves as an organisation ground for the user to construct groupings of images. Groupings of images can be created by right-clicking anywhere on the workspace, which opens a context menu in which the option can be selected. Traditional drag-and-drop techniques allow the user to drag images into (or out of) a group or reposition the group on the workspace. An image can belong to multiple groups simultaneously. Panning and zooming techniques are supported to assist navigation in a large information space. Also, the recommendations will be displayed close to the selected group on the workspace (see centre of workspace in Figure 1). So as not to burden the user, the number of recommended images (set to 10 in this evaluation) is based on the standard cognitive limits of 7 ± 2 [14].

To recapitulate, the query facilities available in the WS interface are: (1) manually constructed queries by providing one or more image examples (QBE), and (2) user-requested recommendations.

2.3 Relevance Feedback Interface - CS

The baseline system is a traditional relevance feedback system, referred to as *CS* (for Checkbox System). Relevance feedback (RF) is an automatic process of improving the initial query based on relevance judgements provided by the user [7]. The process is aimed at relieving the user from having to reformulate the

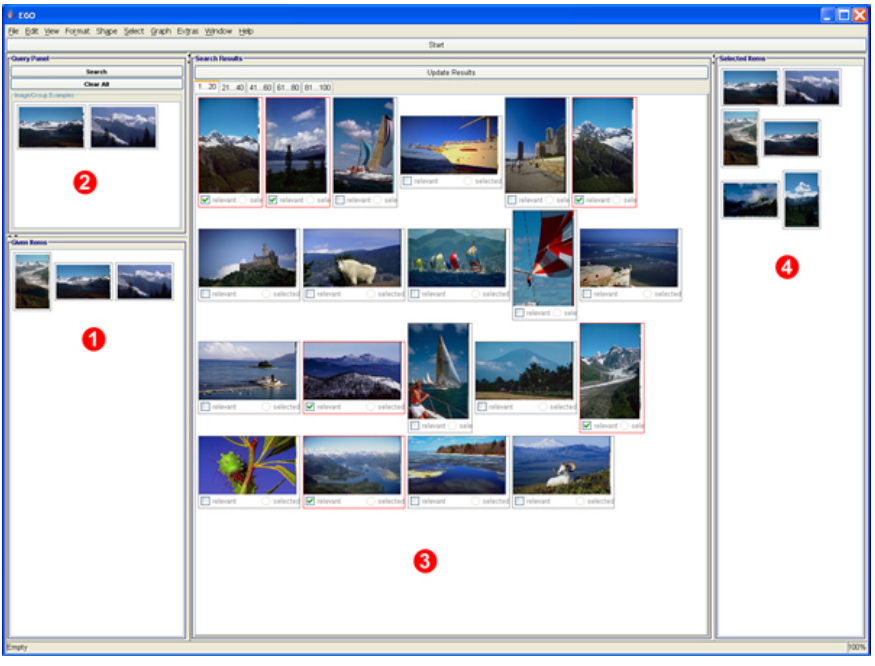

Fig. 2. Annotated CS interface

query in order to improve the retrieval results incrementally. The search becomes more intuitive to the user, since they are only requested to label the returned images as either relevant or not. However, it is still an ongoing research challenge to accurately learn the information need from the user based on a few relevance judgements [15].

Figure 2 shows the CS interface with the following components:

1. Given Items Panel: as above.
2. QBE Panel: as above.
3. Results Panel: As above, but instead of dragging a relevant image onto the workspace the user has the choice of labelling it by selecting a check-box underneath the image. After relevant images have been marked the user can ask the system to update the current search results (based on the feedback provided) by clicking the "Update Results" button in this panel.
4. Selected Items Panel: Any item selected relevant during the course of the search session will be added to this panel. The user can manually delete images from this panel if they change their mind at a later change.

Finally, CS supports two query facilities: (1) QBE as above, and (2) automatic query reformulation by the user feedback provided in the search results (RF).

3 Experimental Methodology

It has been argued that traditional IR evaluation techniques based on precision-recall measures are not suitable for evaluating adaptive systems [16, 17]. Thus in order to evaluate the systems, we used a task-oriented, user-centred approach [18]. We have designed the experiments to be as close to real-life usage as possible: we have chosen participants with a design-related background and have set tasks that are practical and relevant.

In our evaluative study, we adopted a randomised within-subjects design, in which 12 searchers used two systems. The independent variable was system type; two 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 counterbalance the effect of learning from one system to the other, the order of the systems and tasks was rotated according to a Latin square design. For the purpose of the experiment we employed a subset of the Corel collection (CD 1, CD 4, CD 5, and CD 6 of the Corel 1.6M dataset), containing 12800 photographs in total.

3.1 Tasks

In order to place our participants in a real work task scenario, we used a simulated work task situation as conducted in [16]. This scenario allows the users to evolve their information needs in just the same dynamic manner as such needs might be observed to do so in participants' real working lives. A description of the work task scenario and tasks is provided in Figure 3.

Task Scenario
Imagine you are a designer with responsibility for the design of leaflets on various subjects for the Wildlife Conservation (WLC). The leaflets are intended to raise awareness among the general public for endangered species and the preservation of their habitats. These leaflets [...] consisting of a body of text interspersed with up to 4–5 images selected on the basis of their appropriateness to the use to which the leaflets are put.
Category Search Task:
You will be given a leaflet topic from the list overleaf. Your task involves searching for as many images as you are able to find on the given topic, suitable for presentation in the leaflet. In order to perform this task, you have the opportunity to make use of an image retrieval system, the operation of which will be demonstrated to you. You have 10 minutes to attempt this task.
Design Task:
This time, you're asked to select images for a leaflet for WLC presenting the organisation and a selection of their activities (some of WLC's activities are listed overleaf but feel free to consider other topics they might be involved in). Your task is to search for suitable images and then make a pre-selection of 3-5 images for the leaflet. You have 20 minutes to attempt this task.

Fig. 3. Task Description

We created two different tasks: one resembling category search (i.e. users were asked to find as many images as possible from a given topic); and the other resembling an open-ended design task, where they had to search for and make a choice of 3-5 images. The first task was set on both systems, CS and WS, while the latter one was performed on WS only after having completed the category searches. A maximum time was set for all tasks in order to limit the total time spent on the experiment. This was 10 minutes for the category search, and 20 minutes for the design-task.

3.2 Hypotheses

The hypotheses investigated in this study are that the proposed approach for image retrieval and management helps the user to conceptualise their search tasks and to overcome the query formulation problem. The following sub-hypotheses provide more justification:

- Grouping search results on the workspace incites the user to organise results for their search/work task, which in turn helps the user to solve the task. (Organisation as a secondary notation in support of memory/information seeking.)
- The recommendation system helps to overcome the query formulation problem, because it is closer to "real life" search strategies.

In particular we investigate user's performance on WS compared to a standard relevance feedback interface, CS. The latter relies on relevance assessments provided by the user explicitly by marking images from the search results as relevant. Our assumption is that the relevance assessment in CS might be easier and quicker to use, but is less transparent to the user in comparison to creating groups on the workspace in WS, where the user has control over which images belong together. The option of interactively grouping the search results is assumed to be more natural to the user and to lead to a higher level of control.

3.3 Participants

Since we wanted to test the system in a real-life usage scenario, our sample user population consisted of post-graduate design students and young design professionals. Responses to an entry-search questionnaire indicated that our participants could be assumed to have a good understanding of the search and 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.

All participants were in the age group of 20-30 years. There were 9 male participants and 3 female. They had on average 5 years experience in a design-related field (graphic design, architecture, photography). Most people dealt with digital images at least once a day as part of their course or work.

The participants were also asked about their prior experience with search engines and services for searching for images, and image management systems for

organising their own images. Every participant had used an internet image search engine before, whereas only 5 people had used a stock image collection (such as Getty Images, Corbis, Corel). Concerning the organisation of their images, 9 people did not use any management system but just organised their images into folders. The image management systems that were used by the remaining 3 users were ACDSee, Picasa and Extensis Photo Studio.

3.4 Procedure

We met each participant on a separate occasion and adhered to the following procedure:

- an introductory orientation session
- a pre-search questionnaire
- a hand-out of written instructions for the tasks and setting the scenario
- **Part 1:** category search
  - for each system (CS and WS)
    - a training session on the system
    - a search session in which the user interacted with the system (max 10 min)
    - a post-search questionnaire
  - a questionnaire comparing the two systems
- **Part 2:**
  - a search session on WS system (max 20 min)
  - a post-search questionnaire

The total time for one session was 120 min.

4 Results Analysis

There are two objectives of this experiment: (1) to compare the two systems according to their effectiveness and user satisfaction; and (2) to analyse how people make use of the workspace depending on the nature of the tasks. These two parts of the results analysis are expected to shed light on the experimental hypotheses that WS helps users to both conceptualise their tasks better and overcome the difficulties with formulating queries.

4.1 System Comparison

The first objective of the experiment was to compare the two interfaces. It involved two category search tasks, one on each system. The analysis is based on data obtained through questionnaires and usage logs. The questionnaires present a subjective view indicative of the system's acceptability and usability from the users' perspective, while the log data provides a means of judging the task performance objectively. In the questionnaires, we used 5-point semantic differentials, 5-point Likert scales and open-ended questions. Tests (using the non-parametric Wilcoxon Paired-Sample test) for statistical difference will be given where appropriate with $p \leq .05$, unless otherwise stated. The results for the semantic differentials and Likert scales are in the range $[1, 5]$, with 5 representing the best value. $\overline{CS}$ and $\overline{WS}$ denote the means for CS and WS respectively, while $\widetilde{CS}$ and $\widetilde{WS}$ denote the medians.

Task Performance. Data in the usage logs sheds light on how people actually used the system. From this data we can obtain information on the number of relevant images found over the course of the search session. (The ground-truth was obtained by manually labelling relevant images.) Table 1 shows the number of relevant images for each of the tasks and systems. The total number of relevant images varies greatly per task. The level of recall (number of relevant images found over number of total relevant images for the task) attained depends therefore not only on the complexity of the task but also on the number of relevant images available in the system. The tasks were chosen so that Tasks 1-3 represented simple and concrete topics ("mountains", "tigers", "elephants"), while Tasks 4-6 comprised multiple facets ("animals in the snow", "African wildlife", "underwater world"). Looking at the data in Table 1 it can be inferred that users generally performed better in CS independent of the nature of the task.

Table 1. Number of relevant images found and corresponding levels of recall per task

System	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	All
Total #Relevant Images	549	114	103	220	865	402	375.5
#Images found AVG	56.5	14.0	15.25	44.0	38.75	36.75	34.2
#Images found CS	71.5	18.0	18.5	54.5	50.5	34.0	41.2
#Images found WS	41.5	10.0	12.0	33.5	27.0	29.0	25.5
Recall AVG	10.3%	12.3%	14.8%	20.0%	4.5%	7.8%	11.6%
Recall CS	13.0%	15.8%	18.0%	24.8%	5.8%	8.5%	14.3%
Recall WS	7.6%	8.8%	11.7%	15.2%	3.1%	7.2%	8.9%

User Satisfaction. After having completed a task the participants were given a questionnaire about their search experience (post-search questionnaire). Finally, they were asked to compare the two systems in the exit questionnaire. In this section we will analyse the users' opinion on the systems as inferred from the answers provided in the questionnaires.

Post-Search Questionnaire. In the post-search questionnaire people were asked about the task they performed, the images received through the searches, and the system itself.

Task and Search Process. In general, the tasks were considered *clear* and *familiar*, but slightly more *simple* in CS (see Table 2). The search process was considerer slightly more *relaxing* and *easier* in CS, but significantly more *interesting* in WS. However, people tended to agree more with the statement that they had enough time to complete their task in CS: $\overline{CS} = 4.6$, $\widetilde{CS} = 5$ and $\overline{WS} = 4.3$, $\widetilde{WS} = 4$.

Images. The images received through the searches were considered equally *relevant* and *appropriate*, but significantly more *complete* in WS (see Table 4). More people agreed with the statement, that they discovered more aspects of the category than initially anticipated during the search on WS ($\overline{CS} = 2.4$, $\widetilde{CS} = 2$ and $\overline{WS} = 4.4$, $\widetilde{WS} = 5$; $p = 0.02$). On the other hand, people tended to be equally satisfied with their search results in both systems ($\overline{CS} =$

Table 2. Semantic Differentials Results for the Task and Search Process Part

Differential	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	p
clear	4.8	5	4.8	5	-
familiar	3.8	4	3.7	4	-
simple	4.8	5	4.5	5	-
relaxing	4.6	5	3.9	4	-
easy	4.5	5	4.3	5	-
interesting	3.6	4	4.3	4	0.016

Table 3. Semantic Differentials Results for the System Part

Differential	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	p
wonderful	3.7	4	4.1	4	-
satisfying	3.9	4	4.1	4	-
stimulating	3.2	3	3.8	4	0.004
easy	4.6	5	4.1	4	0.031
flexible	2.8	3	3.9	4	0.001
novel	3.1	3	4.2	4	0.016
effective	4.3	4	4.3	4	-

Table 4. Semantic Differential Results for the Images Part

Differential	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	p
relevant	4.2	4	4.2	4	-
appropriate	4.2	4	4.3	4	-
complete	3.3	3	4.1	4	0.027

Table 5. Likert Scale Results for System Part

Statement	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	p
learn to use	4.8	5	4.1	4	0.03
use	4.5	5	4.0	4	-
explore collection	3.3	3	4.3	4	0.03
analyse task	3.1	5	4.5	5	0.02

Table 6. Comparison of system rankings

System	(a) learn	(b) use	(c) effective	(d) liked best
CS	5	5	4	3
WS	3	6	6	8
no difference	4	1	2	1

3.6, $\widetilde{CS} = 4$ and $\overline{WS} = 3.6$, $\widetilde{WS} = 4$). There is no apparent correlation between actual task performance and perceived task performance.

System. The users considered CS significantly more *easy* than WS, while they considered WS to be significantly more *stimulating*, *flexible*, and *novel*. Table 3 shows the results for these differentials.

People found CS significantly easier to *learn to use*, while there was only a marginal difference between *using* them. However, people thought WS helped them to explore the collection better, as well as analyse the task better. The results for the responses to these statements are provided in Table 5.

Exit Questionnaire. After having completed both category search tasks having used both systems, the users were asked to determine the system that was (a) easiest to learn to use, (b) easiest to use, (c) most effective, and (d) they liked best overall. Table 6 shows the users' preferences of systems for each of the statements. It shows that, while it is easier to learn to use CS, the majority of people preferred WS and found it more effective.

In open-ended questions, the participants were invited to give their opinion on what they liked or disliked about each system. The advantages listed

for CS were that it was fast, efficient and easy to use. Its disadvantages included that the users felt they did not have enough control over the search and that its interface was less intuitive. In WS, people liked the ability to plan their searches by organising the results into groups, and the overview they had of the results and searches that the organisation brought along. In addition, the system's flexibility and more control options were noted as advantages. The disadvantages were mainly concerned with the poor quality of the recommendations and that the handling of groups was sometimes cumbersome. Both of these issues are not inherent in the interaction paradigm of the proposed system itself, and can consequently be improved or even avoided in the future. The recommendation quality can be improved by a better choice of visual features and also by recommendations based on other people's groupings. The handling of the groups and images within groups is a matter of programming.

4.2 Task Analysis

The second objective of the study is to judge the usefulness of the workspace to help the user to conceptualise their task. In order to find out how people make use of the groupings and organise their workspace, we have created two different task scenarios in the experiment: the category search scenario and the design task scenario. The former (set on both WS and CS) aims at maximising recall, while the latter aims at finding a selection of good quality images that work well together (only on WS). By analysing the number of groups created and the average number of images per group for the various tasks, we can identify how these numbers relate to task complexity.

Unfortunately, we cannot present the full analysis in this paper. Please refer to [19]. To summarise, we found a correlation between the number of groups created and the complexity of the task set. Further, responses in the questionnaires showed that the management of search results was deemed more helpful in the design scenario, which is more flexible and open to interpretation than the category search scenario. In the category search scenario, the usefulness of the organisation also depended on the complexity of the task: the more facets the task comprised, the more useful the workspace was considered. This strong dependency between both the number of groups created and the users' perception of the workspace's usefulness, led us to the conclusion that our approach indeed helps in conceptualising the task.

5 Discussion

By analysing users's behaviour in different task scenarios, we have been able to show that the grouping facility was used to reflect the various task facets, and therefore helped to conceptualise tasks. On the other hand, it is more difficult to draw a definite conclusion on the second hypothesis, namely that our approach helps to overcome the query formulation problem. The responses in the questionnaires suggest that the search process is more interesting in WS, the system

helped them to discover more aspects of the task, and found it more stimulating, flexible and novel. In general, they preferred WS over CS and found it more effective for the task. The participants particularly liked the ability to plan their searches and organise their results. In comparison, they considered they were lacking control over their searches in CS. However, the actual task performance does not reflect the users' perception. The number of relevant images found per task were generally higher in CS than in WS. Based on the analysis of the questionnaire data above, the reason for this is that the selection of relevant images is much faster than the dragging of images. Also, the users spent time on creating groups of images and moving images between groups in the WS system. Since we have set a maximum time limit, the number of images found was generally higher in CS, where the user was not "distracted" by managing their search results.

In addition, the failure of the recommendation system has most probably contributed to these results. Analysing the users' comments, we could identify that many people thought the recommendation system would potentially have been a very useful feature, but was not employed due to its inability to recommend relevant images. Our initial hypothesis, namely that the recommendation system helped to overcome the query formulation problem, could not be verified directly. On the other hand, when analysing the way the users manually created the queries, we could observe an interesting pattern. They usually started off with a small number of example images (from the given items, and some initial results). Once they had created a group on the workspace that contained a number of relevant images, they used the whole group in the QBE search to find similar images to the *group*. We assume that, had the recommendation system worked better, users would have used the recommendations in that case. However, since this was not the case, they had to resort to the manual facility of finding more similar images for the group.

In conclusion, the difference in performance can be attributed to the additional effort – both physical (slower selection process) and cognitive – required in WS. While the users commented on the additional physical effort, they did not perceive the additional cognitive effort as negative. On the contrary, they thought the organisation to be supportive for solving their tasks as well as potentially beneficial for others to use in the future.

6 Conclusion and Future Work

We have presented a user study comparing the proposed EGO system to an image retrieval system with relevance feedback capabilities. While the performance in a category search task was generally higher in the relevance feedback system, the proposed system led to a higher user satisfaction. We identified possible reasons for the differences in task performance: the time restriction was limiting and the recommendation system's performance was not good enough. Still, the participants preferred our system, because it allowed them to organise their search results and hence conceptualise the task better.

Since we have encountered differences in the perceived usefulness of the grouping facility depending on the task nature, we believe the interface should have a way to be tailored to these contrasting requirements to adapt to its users. In the future, the user should be assisted in determining task aspects and create groups (semi-) automatically. For a multi-aspect task, we could then group results into the various aspects and present recommendations for each group.

Moreover, a more sophisticated active learning approach, such as the one proposed in [20], could help to improve the recommendations based on the visual features of the images. In addition, the recommendations should also incorporate information from the groups created by the users. This can be used to learn associations between images, and, when combined with the visual similarity, lead to not only more accurate recommendations but also to personalised recommendations. This is the case, since similarity between images would then be based on semantic concepts as defined by the users. We would also like to investigate EGO in a collaborative context. By placing the resulting groups of images on a workspace, the user creates traces of their activities. These traces could be used in a collaborative environment in two ways: first, the system can use the groups created by various users to learn general and personal associations between images; and second, by inspecting someone else's workspace one can retrace their activities.

The real benefits of such a management system will only have an effect if it is used over a longer period of time. The organisation of the collection created over time is an important clue for the system to learn and improve its recommendations over time. The interaction data collected in this study will therefore be useful in follow-up evaluations requiring a long-time involvement of the user.

References

1. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
2. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. Computer **28** (1995) 23–32
3. Lim, J.H.: Learnable visual keywords for image classification. In: Proc. of the ACM Int. Conf. on Digital Libraries (DL-99), ACM Press (1999) 139–145
4. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of the Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR03). (2003) 119–126
5. ter Hofstede, A.H.M., Proper, H.A., van der Weide, T.P.: Query formulation as an information retrieval problem. The Computer Journal **39** (1996) 255–274
6. Urban, J., Jose, J.M.: EGO: A personalised multimedia management and retrieval tool. Int. Journal of Intelligent Systems (IJIS), Special Issue on 'Intelligent Multimedia Retrieval' (2005) to appear.
7. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol. **8** (1998) 644–655

8. Stricker, M., Orengo, M.: Similarity of color images. In: Proc. of the SPIE: Storage and Retrieval for Image and Video Databases. Volume 2420. (1995) 381–392
9. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. 2nd edn. Brooks and Cole Publishing (1998)
10. Hu, M.K.: Visual pattern recognition by moment invariants. IEEE Trans. Information Theory **8** (1962) 179–187
11. Rui, Y., Huang, T.S.: Optimizing learning in image retrieval. In: IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR-00) (2000) 236–245
12. Porkaew, K., Chakrabarti, K., Mehrotra, S.: Query refinement for multimedia similarity retrieval in MARS. In: Proc. of the ACM Int. Conf. on Multimedia (1999) 235–238
13. Urban, J., Jose, J.M.: Evidence combination for multi-point query learning in content-based image retrieval. In: Proc. of the IEEE 6th Int. Symposium on Multimedia Software Engineering (ISMSE'04) (2004) 583–586
14. Miller, G.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review **63** (1956) 81–97
15. Zhou, X.S., Huang, T.: Relevance feedback in image retrieval: A comprehensive review. ACM Multimedia Systems Journal **8** (2003) 536–544
16. Jose, J.M., Furner, J., Harper, D.J.: Spatial querying for image retrieval: A user-oriented evaluation. In: Proc. of the Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98), ACM Press (1998) 232–240
17. de Vries, A.P.: The role of evaluation in the development of content-based retrieval techniques. Technical Report TR-CTIT-00-19, Centre for Telematics and Information Technology (2000)
18. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham, London (1992)
19. Urban, J., Jose, J.M.: Exploring results organisation for image searching. In: Proc. of INTERACT 2005, Springer (2005) to appear.
20. Jin, R., Chai, J.Y., Si, L.: Effectiv automatic image annotation via a coherent language model and active learning. In: Proc. of the ACM Int. Conf. on Multimedia, ACM Press (2004) 892–899