Multimedia Systems (2007) 12:355–373
DOI 10.1007/s00530-006-0051-z

REGULAR PAPER

Evaluating a workspace's usefulness for image retrieval

Jana Urban · Joemon M. Jose

Published online: 16 September 2006

Abstract Image searching is a creative process. We have proposed a novel image retrieval system that supports creative search sessions by allowing the user to organise their search results on a workspace. The workspace's usefulness is evaluated in a task-oriented and user-centred comparative experiment, involving design professionals and several types of realistic search tasks. In particular, we focus on its effect on task conceptualisation and query formulation. A traditional relevance feedback system serves as a baseline. The results of this study show that the workspace is more useful in terms of both of the above aspects and that the proposed approach leads to a more effective and enjoyable search experience. This paper also highlights the influence of tasks on the users' search and organisation strategy.

Keywords User evaluation · Image retrieval · Relevance feedback · Workspace · Recommendation system

1 Introduction

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 this lack of acceptance: 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*) [1].

Our objective is to find a solution to these problems by supporting an alternative search strategy. We have designed a system, EGO, that combines the search and the management process [2]. This is accomplished by introducing a workspace alongside a recommendation system. 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.

We have designed a user experiment to evaluate the effectiveness of our approach for solving real-life image search tasks. We compare EGO's performance to that of a traditional relevance feedback system. 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. Our aim is to collect evidence on the systems' effectiveness as perceived by the users and in particular, we will focus on the workspace's usefulness. By observing and analysing the user's organisation strategy we will answer the following questions: How was the workspace used? What influence did the task have on this? More importantly, however, we would like to determine the workspace's role in helping the user to both conceptualise

J. Urban (✉) · J. M. Jose
Department of Computing Science,
University of Glasgow,
17 Lilybank Gardens, G12 8RZ, Glasgow, UK
e-mail: jana@dcs.gla.ac.uk

J. M. Jose
e-mail: jj@dcs.gla.ac.uk

their search tasks and overcome the query formulation problem.

The experiment was completed in two stages. The underlying experimental methodology is described in Sect. 4. Experiment 1 involved 12 participants using the two systems for category search tasks and a design task. The results analysis is presented in Sect. 5. In this experiment, we only studied two different tasks from which the design task was only performed on the workspace system. So to be able to further study the effect of task on searching and organisation behaviour, we needed to investigate a larger variety of tasks. Therefore, a second experiment was conducted with new tasks and 12 new participants, thus allowing us to analyse the usefulness of the system in a wider context. Sect. 6 summarises our findings. We provide the combined results of Experiments 1 and 2 in Sect. 7 and a summarising discussion in Sect. 8. Finally, Sect. 9 concludes the paper.

2 Motivation

Image retrieval systems are tainted by problems caused by the interaction with typical CBIR user interfaces mainly due to the semantic gap and the query formulation problem. It has become apparent that the *user* plays a very – if not *the* most – important role. Without the users' knowledge of the world and their superior visual system, CBIR system capabilities are rather limited. Moreover, user satisfaction greatly depends on subjective judgements of image contents as well as relevance. It is impossible to accommodate the huge diversity of users, yet systems can adjust to individual users by learning their preferences.

From the user's perspective, however, searching for and performing a selection of images is usually embedded in other tasks, and thus it is at least equally important to understand and capture the work flow [3,4]. Therefore, a solution to accommodate the needs of users must be flexible, should support multiple tasks, and allow exchanges or even seamless integration with other applications used for the work tasks. Moreover, the search process often takes place in a collaborative context, in which people work together, are inspired and learn from each other's activities.

What is needed is a *holistic view* on personal image organisation and retrieval. A "retrieval in context system" offers a great opportunity for learning, adaptation and personalisation, which can overcome the problem of detecting the usage context dependent meaning of images. These considerations have led us to the design of EGO (Effective Group Organisation). 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 same problems render image retrieval systems particularly difficult to evaluate. To date, there still does not exist a common testbed despite several efforts to this end (e.g. the Benchathlon network [5] and more recently ImageCLEF [6]). What makes creating a testbed so challenging is the lack of objective measures for realistic image search tasks. People have employed category search and target search tasks, where the set of relevant images can be determined beforehand and hence traditional precision and recall measures [7] can be used. However, image searching is an inherently creative activity. Our target user population is expected to use our system for design-related work tasks. In these scenarios, it is seldom the case that an image retrieval system is consulted to search for such a clearly defined set of images. On the contrary, the underlying information need is typically vague, and the result set is fuzzy.

For these reasons, we have adopted a user-centric, task-oriented experimental methodology. We have devised several design-oriented tasks and asked design professionals to participate in order to create a realistic search experience. Each task description is accompanied by a scenario, which describes a simulated work task [8]. The simulated work task situation is aimed at re-creating tasks from an individual's real working life. This allows the users to develop their own interpretation of the task and use their own judgement for choosing relevant images. This way, we can study how information needs evolve and what influence the interface has on their search and organisation strategy.

3 The EGO system

3.1 Retrieval system

The underlying retrieval system has been described in [2,9]. Images are represented by a set of low-level visual features and modelled according to the hierarchical object model [10]. The distance between an object in

the database and a given query representation is computed in two steps: computing the individual feature distances by the generalised Euclidean distance; then combining the individual distances linearly with a set of feature weights. The relevance feedback algorithm is implemented by an optimised framework for updating the retrieval parameters as proposed in [10]. It attempts to learn the best query representation and feature weighting for a selected group of images (positive training samples).

3.2 The interface

The EGO interface depicted in Fig. 1 comprises the following components:

1. *Query panel*: This provides a basic query facility to search the database by allowing the user to compose a search request by entering search terms or adding example images to the query-by-example (QBE) panel provided here. Clicking on the "Search" button in this panel will issue a search.
2. *Results panel*: The search results from a query constructed in the Query 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.
3. *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 can be created by right-clicking anywhere on the workspace, which opens a context menu in which the option can be selected or alternatively using a button located in the toolbar on the top of the workspace. 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 are displayed close to the selected group on the workspace (see centre of workspace in Fig. 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 [11].
4. *Group results panel*: For each query or recommendations issued the existing groups will be ranked in order of similarity to the current query/group and the five top matching groups will be displayed in this panel. Each returned group contains a link to the original group on the workspace.

4 Experimental methodology

It has been argued that traditional IR evaluation techniques based on precision-recall measures are not suitable for evaluating adaptive systems [8,12]. Hence, we used a task-oriented, user-centred approach [12]. 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.

We employed a subset of the Corel collection (CD 1, CD 4, CD 5, and CD 6 of the Corel 1.6M dataset), containing 12,800 photographs in total. Twenty-four searchers used two systems in a randomised within-subjects design.

A within-subjects design is an experiment in which the same set of dependent variables is measured repeatedly on the same participant under different "treatments" (levels of independent variables). In our case, the treatments are system type and task type. The dependent variables are the responses from the questionnaires and other data collected from usage logs. The advantage of a within-subjects design is that effects due to the disposition of participants are minimised. This is beneficial because the variability in measurements is more likely due to differences among conditions than to behavioural differences between participants. There is one major weakness of this type of design: the learning effect, as participants' behaviour in one condition will affect their behaviour in another.

To counterbalance the effect of learning, the order of the systems and tasks was rotated according to a Latin-square design. A Latin square is an $n \times n$ table filled with n different conditions in such a way that each condition occurs exactly once in each row and exactly once in each column. Figure 2a shows an example Latin square for two conditions. In this case, participants are randomly assigned to groups of equal size: Group 1 is given condition A followed by condition B, while Group 2 is given condition B followed by condition A. Figure 2b and c shows the Latin squares if three or four conditions are tested.

4.1 The interfaces

1. *Workspace interface – WS*: The interface used in the evaluation is a simplified version of the EGO interface. 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, this interface is referred to as the Workspace Interface (WS). The query

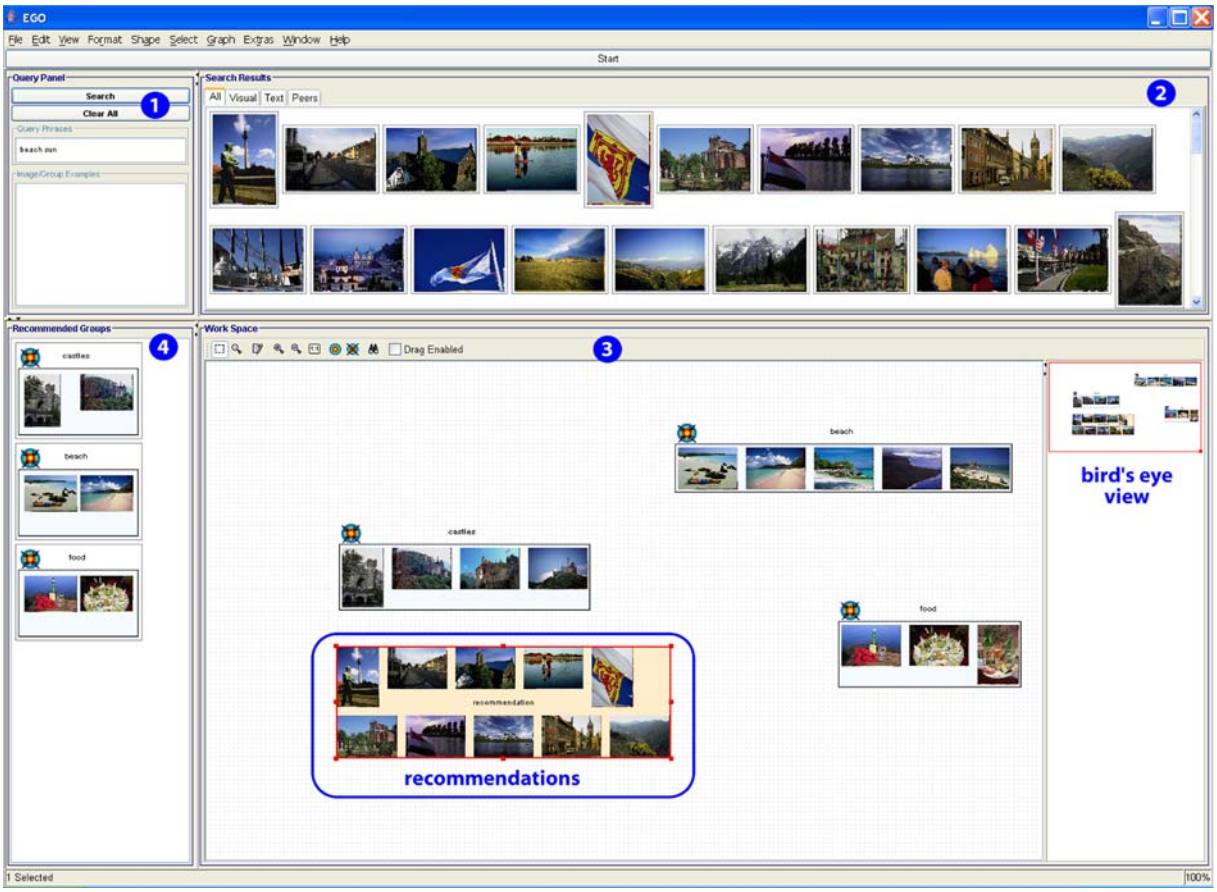

Fig. 1 Annotated EGO interface

a)

A	B
B	A

b)

A	B	C
C	A	B
B	C	A

c)

A	B	C	D
D	A	B	C
C	D	A	B
B	C	D	A

Fig. 2 Example Latin squares

facilities available in the WS interface used in Experiment 1 are: (1) manually constructed queries by providing one or more image examples (QBE), and (2) user-requested recommendations (REC). The experimental interface also comprises a Given items panel, which contains a selection of images (three per task) provided for illustration purposes and can be used to bootstrap the search. This panel replaces the Group results panel in Fig. 1.

2. *Relevance feedback interface – CS*: The baseline system is a traditional relevance feedback system, referred to as CS (for Checkbox System). Figure 3 shows the CS interface with the following components:
   1. *Given items panel*: As above.
   2. *Query panel*: As in Sect. 3.2.
   3. *Results panel*: As in Sect. 3.2, but instead of dragging a relevant image onto the workspace the user has the choice of labelling it by selecting a checkbox 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*: All items selected relevant during the course of the search session are added to this panel. The user can manually delete images if they change their mind at a later change.

To summarise, the look-and-feel of the interface is similar to WS (without the workspace facility). Finally, CS supports two query facilities: (1) manual queries as above (QBE), and (2) automatic query reformulation by the feedback provided in the search results (RF).

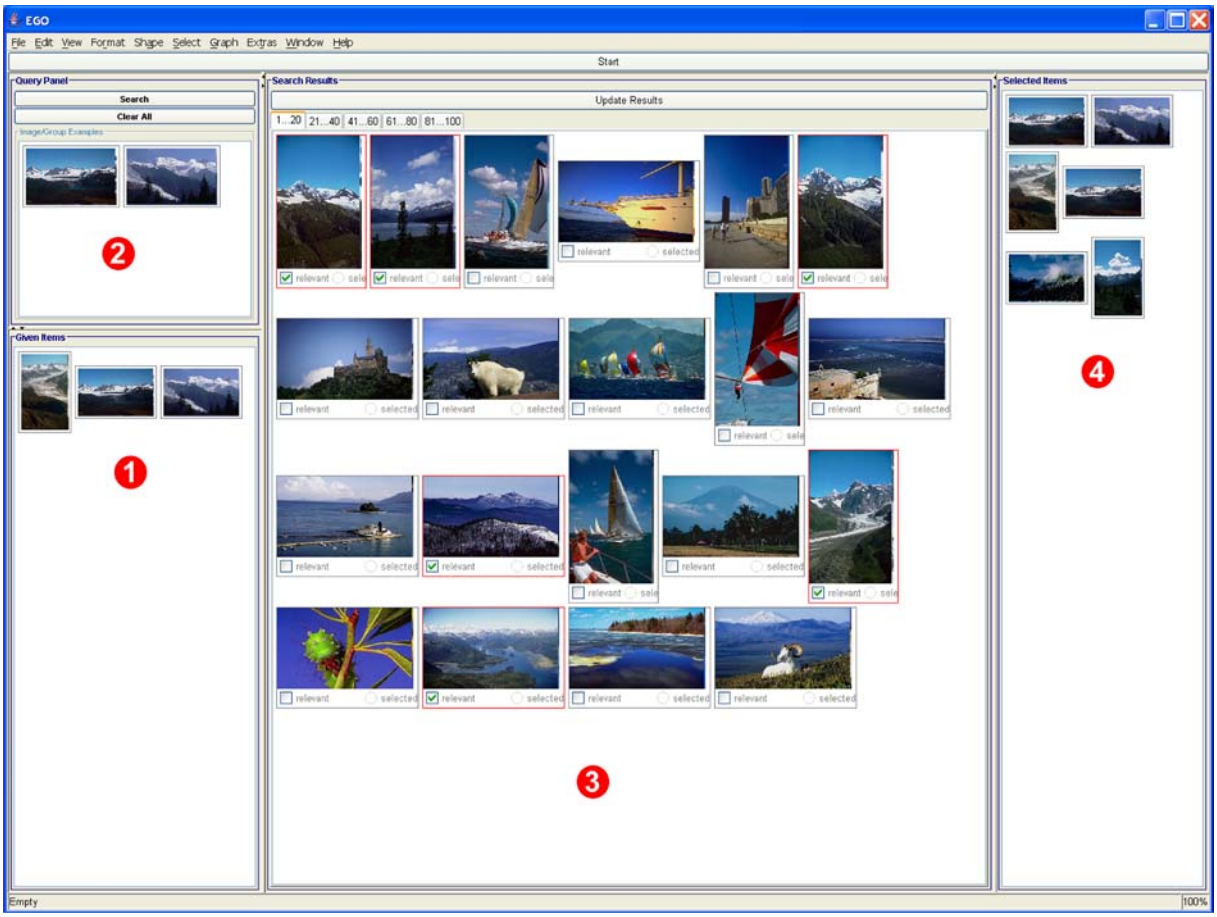

Fig. 3 Annotated CS interface used in Experiment 1

4.2 Participants

Our sample user population consisted of post-graduate design students and young design professionals. Responses to an entry questionnaire indicated that our participants could be assumed to have a good understanding of the search and design task we were to set them. We could also safely assume that they had no prior knowledge of the experimental systems. There were 24 participants in total: 16 male and 8 female. The age range was 20–50 with an average age of 27 years. They had on average 5 years experience in a design-related field (graphic design, architecture, and photography). Most people dealt with digital images at least once a day as part of their course or work.

4.3 Tasks

We used a simulated work task situation as conducted in [8]. This scenario allows for evolving information needs in just the same dynamic manner as might be observed in an individual's real working life. A description of the work task scenario and tasks is provided in Fig. 4. The tasks were:

A/B. In the *category search* scenario users were asked to find as many images as possible from a given topic. The topics in Task A represent simple and concrete topics ("mountains", "tigers", "elephants"), while the topics in Task B comprise multiple facets ("animals in the snow", "African wildlife", "underwater world").

C. The *design task* resembles an open-ended design task, where the participants had to search for and make a choice of 3–5 images.

4.4 Hypotheses

As investigating the workspace's usefulness is a high-level goal, the experimental hypothesis has been broken up into the following more manageable sub-hypotheses:

1. The addition of a workspace leads to a more effective system and increased user satisfaction.

Task Scenario
Imagine you are a designer with responsibility for the design of leaets on various subjects for the Wildlife Conservation (WLC). The leaets are intended to raise awareness among the general public for endangered species and the preservation of their habitats. These leaets [...] 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 leaets are put.
Category Search Task (Tasks A and B):
You will be given a leaet topic from the list overleaf. Your task involves searching for as many images as you are able to nd on the given topic, suitable for presentation in the leaet. 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 (Task C):
This time, you're asked to select images for a leaet 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 leaet. You have 20 minutes to attempt this task.

Fig. 4 Task description for Experiment 1

2. The workspace helps users to conceptualise their tasks better.
3. The grouping and recommendations help to overcome the query formulation problem.

4.5 Procedure

Our experiment started with an introductory orientation session and a pre-search questionnaire. The actual search tasks were divided into two parts: the category search part and the design part. For the first part, the participants performed one category search task on each system (one topic of Task A on the first system and one of Task B for the second system). Each search session (max 10 min[1]) was preceded by a training session on the system, and followed by a post-search questionnaire. After having completed the two search sessions, the participants were asked to complete an exit questionnaire comparing the two systems. For the second part, the participants were asked to perform the design task on WS (max 20 min), followed by a post-search questionnaire. The total time for one experiment was 120 min.

4.6 Data capture

1. *Questionnaires*: The questionnaires elicit people's opinion on the tasks performed, the images found during the search session, the usability of the systems and their satisfaction with their task performance. Their opinion was captured on five-point semantic differentials, five-point Likert scales and open-ended questions. The results for the semantic differentials and Likert scales are in the range $[1,5]$, with 5 representing the best value. In the results analysis, statistically significant differences are provided where appropriate with $P \leq 0.05$ using the two-tailed version of the non-parametric Wilcoxon Paired Sample test. $\overline{\text{CS}}$ and $\overline{\text{WS}}$ denote the means for CS and WS, respectively, while $\widetilde{\text{CS}}$ and $\widetilde{\text{WS}}$ denote their medians.
2. *Usage logs*: The data logged include total session time, images selected during the search, types and number of queries issued. These results are analysed and summarised to reflect the users' performance and required effort to complete the tasks.

5 Experiment 1

The first stage of the experiment involved 12 participants using the two experimental systems on the tasks described above. 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 in order to determine its effect on task conceptualisation.

5.1 System comparison

The comparison is based on the category search tasks, involving 24 searches in total (one topic per system per user). The questionnaire results present a subjective view indicative of the system's acceptability and usability, while the log data provides a means of judging task performance objectively.

1. *Task performance*: From the usage logs we can obtain information on the total number of relevant images found for the category search tasks.[2] Table 1 shows the number of relevant images for each of the topics 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 topic) attained depends therefore not only on the complexity of the task but also on the number of relevant images available in the system. Users generally performed better on CS independent of the nature of the task. Yet, the questionnaire analysis below suggests that there was a stronger focus in

[1] A maximum time was set for all tasks in order to limit the total time spent on the experiment.

[2] The ground-truth was obtained by manually labelling relevant images in the collection for each topic.

Table 1 Number of relevant images found and corresponding levels of recall per category search topic

	Task A			Task B			AVG
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
Total #relevant	549	114	103	220	865	402	375.5
#Rel AVG	56.5	14.0	15.25	44.0	38.75	36.75	34.2
#Rel CS	71.5	18.0	18.5	54.5	50.5	34.0	41.2
#Rel 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%

WS to find *appropriate* images for the leaflet, facilitated by better tools for exploring the task and the collection.

2. *User satisfaction*: After having completed a task the participants were given a post-search questionnaire about their search experience. Finally, they were asked to compare the two systems in the exit questionnaire. In this section we analyse the users' opinion on the systems as inferred from the answers provided in these questionnaires.
   1. *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*: The first part of the post-search questionnaire covered the user's perception of task complexity. The tasks were rated according to the five-point semantic differentials: clear, easy (vs. difficult), simple (vs. complex) and familiar. The results are shown in Table 2 (scores from 1 to 5, higher = better). There are no significant differences on any of the differentials and all scores are well above 3, showing that the users generally considered the tasks to be *clear*, *easy*, *simple*, and *familiar*. However, the tasks were considered slightly *easier* and *simpler* on CS. Note that their perception depends on the users' overall search experience on a particular system, since these responses are received in the post-search questionnaire.
      - *Search process*: The users were asked to rate the search process according to the five-point semantic differentials: easy (vs. stressful), interesting (vs. boring), and restful (vs. tiring). The search process was considered slightly more *relaxing* and *easier* on CS, but significantly more *interesting* on WS. However, people tended to agree more with the statement that they had enough time to complete their task on CS: $\overline{\text{CS}} = 4.6$, $\widetilde{\text{CS}} = 5$ and $\overline{\text{WS}} = 4.3$, $\widetilde{\text{WS}} = 4$.
      - *Images*: The retrieved images were considered equally *relevant*, but slightly more *appropriate* and significantly more *complete* on WS (see Table 2).
      - *System*: The users considered CS significantly *easier* 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. By contrast, 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 4.
   2. *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. They could choose between WS, CS, and no difference. It turned out that, while it is easier to learn to use CS (42% for CS and 25% for WS), people did not have a problem using WS

Table 2 Semantic differential results for the task, search process, and images parts

Differential	$\overline{\text{CS}}$	$\widetilde{\text{CS}}$	$\overline{\text{WS}}$	$\widetilde{\text{WS}}$	*P*
Task					
Clear	4.8	5	4.8	5	–
Easy	4.5	5	4.3	5	–
Simple	4.8	5	4.5	5	–
Familiar	3.8	4	3.7	4	–
Search					
Relaxing	**4.6**	5	3.9	4	–
Interesting	3.6	4	**4.3**	4	0.02
Restful	3.8	4	3.7	4	–
Images					
Relevant	4.2	4	4.2	4	–
Appropriate	4.2	4	4.3	4	–
Complete	3.3	3	**4.1**	4	0.03

Table 3 Semantic differential results for the System and Interaction Parts

Differential	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	P
System					
Wonderful	3.7	4	4.1	4	–
Satisfying	3.9	4	4.1	4	–
Stimulating	3.2	3	**3.8**	4	0.01
Easy	**4.6**	5	4.1	4	0.03
Flexible	2.8	3	**3.9**	4	0.01
Novel	3.1	3	**4.2**	4	0.02
Effective	4.3	4	4.3	4	–
Inter					
In control	4.3	4	4.2	4	–
Comfortable	4.4	5	4.6	5	–
Confident	4.3	4	4.4	5	–

Table 4 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 col.	3.3	3	**4.3**	4	0.03
Analyse task	3.1	5	**4.5**	5	0.02

(CS: 42%; WS: 50%), and the majority of people thought it was more effective (50% compared to 33% for CS) and preferred WS (67%).

5.2 Organisation analysis and information need development

The second objective of the study is to judge the workspace's usefulness in helping the user to conceptualise their task. We provide a detailed analysis for all tasks based on both experiments in Sect. 7.2.

To summarise, we found a correlation between the complexity of the task and the number of groups created (1.2, 4.3, and 4.4 for Tasks A, B, and C, respectively). 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.

We have also analysed the nature of the information need and compared how each of the systems supported the user in either fulfilling or evolving their needs. The responses suggest the participants had a clearer idea of the images relevant for the task in the category search scenario (average 4.4), compared to the design scenario (3.7). While users' initial idea did not vary much across the systems for the category search tasks ($\overline{CS} = 4.2$, $\widetilde{CS} = 4$ and $\overline{WS} = 4.4$, $\widetilde{WS} = 4$), WS still helped more to develop their need: The users detected more aspects of the category than initially anticipated on WS ($\overline{CS} = 2.4$, $\widetilde{CS} = 2$ and $\overline{WS} = 4.4$, $\widetilde{WS} = 5$; $p = 0.02$). This was especially true for the multi-faceted topics ($\overline{CS} = 2.7$ and $\overline{WS} = 4.7$).

On the other hand, people tended to be equally satisfied with their search results in both systems ($\overline{CS} = 3.6$, $\widetilde{CS} = 4$ and $\overline{WS} = 3.6$, $\widetilde{WS} = 4$). There is no apparent correlation between actual task performance and perceived task performance. This shows that people had other performance criteria apart from finding as many images as possible. As mentioned above, we suspect that the users of WS, by providing better tools for exploring and analysing the retrieved images, concentrated more on selecting images of good quality. This can be explained by the fact that the task description not only asked for finding as many images as possible, but also had the additional qualifying statement "*suitable for presentation in the leaflet*".

5.3 Discussion

By analysing users' 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, the conclusion on the second hypothesis, namely that our approach leads to a more effective and usable interface, was ambiguous.

The responses in the questionnaires suggest that the participants were more satisfied with their overall search experience on WS and that it was at least as effective. By contrast, the actual task performance does not reflect the users' perception. The number of relevant images found per task were generally higher on CS than on WS. Based on the analysis of the questionnaire data, 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. There is an additional cognitive effort underlying these activities on WS. The users spent more time thinking about task aspects and groups to create, as well as about which images are appropriate for the leaflet. Since we have set a maximum time limit, the performance was better on CS where the user was not "distracted" by managing

their search results. On top of this, the task description was found to be ambivalent. We suspect that people had a slightly different objective in WS, which supported a more selective search strategy.

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 useful feature, but was not employed due to its inability to recommend relevant images. The main problem was that only the top ten recommendations were accessible, whereas in CS the top 100 images were shown. The overall hypothesis underlying this work, namely that the recommendation system helps to overcome the query formulation problem, could not be verified directly. On the other hand, when analysing the way the users manually created 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 several 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 instead of the QBE search. 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. Given these results, a second phase of the experiment was designed incorporating a number of new tasks. This would allow us to study the effect of tasks on searching and organisation behaviour further and hopefully lead to more insights on the query formulation process in both systems.

6 Experiment 2

After the results obtained from the first set of participants, we have improved the system by taking into account the lessons learnt. The changes made were:

- The recommendation system in WS was not used to its full potential, due to its inability to recommend relevant images. This has been addressed in two ways. First, instead of just showing the top ten recommendations on the workspace, the Results panel now also shows the complete results (limited to 100 images). Second, a textual search facility has been introduced, because the visual features seemed not sufficient to solve more abstract tasks providing a more realistic search experience. Textual annotations obtained from [13] were incorporated and implemented according to the vector-space model [14]. Visual and textual features are combined using a rank-based list aggregation method [9].
- The retrieval mechanism was further improved by allowing negative feedback, as people complained about the inability to continue a search when the majority of returned images were irrelevant. Since incorporating negative feedback is a difficult endeavour [15], we have opted for a quick and safe approach: irrelevant images are added to a negative filter excluding them to be returned for the same search. It was straightforward to implement this in CS where negative feedback can easily be provided explicitly. In WS however, we have chosen an implicit feedback strategy, whereby an image is automatically added to a negative filter for a group when it has been ignored (i.e. not dragged into this group) after having been returned three times amongst the top ten recommendations.
- A new set of tasks has been introduced. We felt that more tasks were needed in order to draw definitive conclusions on the workspace's usefulness in helping to conceptualise tasks. In addition, after having questioned design professionals about their "usual" kind of work and search tasks it became apparent that they rarely have to perform an exhaustive search on a specific topic as is required in the category search task. Therefore, a greater emphasis was placed on creativity and realism when devising the new set of tasks.
- Finally, no time limit was set on the tasks, addressing this problem in the previous setup and supporting an even more creative search session.

With this improved evaluation set-up, Experiment 2 should help clarify the validity of the experimental hypotheses.

6.1 Changes to the experimental methodology

The following changes were made to the experimental methodology described in Sect. 4: slightly modified

interfaces, a new set of tasks, and a slightly different procedure.

1. *The interfaces*: The interfaces described in Sect. 4.1 were slightly modified. The query panel has an additional text box for a user to enter a set of keywords to use in a search. Since the keyword search provides an adequate solution to bootstrap a search session, we no longer needed to provide a set of given items. Hence, the given items panel is no longer present in the interfaces. WS in Experiment 2 has the same panel layout as the screenshot in Fig. 1. The results panel in the new CS interface is modified to allow negative feedback: the checkbox for marking relevant images is replaced by three combo boxes to mark images as one of relevant, irrelevant, or neutral (see Fig. 5).
2. *Tasks*
   D. In the *theme search task*, a theme was illustrated by three example images and the task involved searching for and selecting *one further image* complementing this set (see Fig. 6; note that the textual description was not shown to the users).
   E. The *illustration task* involved illustrating a piece of text for publication on the WWW or an advertising slogan with *three images*. There were four topics in total from which the participants had to choose two (one on each system). One example scenario and task description is provided in Fig. 7.
   F. In the *abstract search task* people were asked to select *at least one image* representing a given abstract topic. The simulated work task situation prescribed to select an image for a photo competition.
3. *Procedure*: A total of 12 users participated, none of whom had taken part in the first experiment or used the systems before. Each participant performed four search sessions, completing two tasks with a different topic per system. Each system was used twice. Tasks and systems were rotated according to a Latin-square design in order to compensate the learning bias. The procedure involved a pre-search questionnaire, the four search sessions followed by a post-search questionnaire each time, and finally an exit questionnaire/interview comparing the systems. A search session was preceded by a training session if the system was used for the first time. The whole procedure lasted approximately 2 h.

Fig. 5 Relevance feedback in CS

6.2 Results analysis

In the results analysis, the systems are first compared according to (a) their effectiveness, and (b) user satisfaction. Finally, the users' organisations of images on the workspace are analysed and related back to the task that was performed and the nature of the users' underlying information needs. The analysis is based on 48 searches in total, 12 users performing four searches each.

1. *Effectiveness*: The systems' effectiveness is investigated both objectively and subjectively: from the perspective of the required effort as determined from the usage logs and from the perspective of the participants.
   1. *User effort*: Due to a lack of objective performance measures for the tasks in Experiment 2, we provide an analysis of the number of images selected per task and the amount of user effort required to select them. These include: total search time and number of queries issued. People can issue either manual queries – constructed in textual form, by providing image examples or a combination of both – or relevance feedback queries. The latter correspond to relevance feedback iterations in CS or group recommendations in WS.
   Table 5 reveals that less queries were issued and more images were selected on WS. In particular, more *relevance feedback* queries were requested on WS, while more *manual queries* were constructed on CS (with the exception of Task F). The RF queries were particularly useful for Tasks E and F. The search session lasted on average longer on WS. By contrast, Task E stands out for being completed in less time on WS (with a difference of about 4 min) but still achieving a slightly larger selection of images in the end. Again, this indicates that WS is particularly useful for design-oriented tasks. We also observed that the selected images on WS were of better quality. Subjectively, the participants felt that the images they had retrieved on WS were more *complete* ($\overline{\text{CS}} = 3.5$, $\overline{\text{WS}} = 3.8$) and slightly more *relevant* ($\overline{\text{CS}} = 4.0$, $\overline{\text{WS}} = 4.1$). Also, they per-

Fig. 6 Image sets for task D topics

ceived their performance more successful (see below). In addition, we asked people to judge the relative quality of the result sets obtained in the experiment to quantify this observation. We randomly selected 16 pairs of result sets per task (48 in total), where one set was retrieved with WS and the other with CS. The participants were given a copy of the original task description and a pair of result sets obtained for this task. They were then asked to (a) select one overall image from the two sets, which – in their opinion – was most suitable for the task; (b) cross out any images they thought were not relevant for the task; and (c) state which set they preferred overall. Sixteen people, judging on average three different pairs each, took part in this study. Out of the 48 pairs, there were 36 preferences for the sets obtained with WS and only 12 for the CS sets. Only on three occasions, people picked the best image from the set they did not prefer, and on one occasion the best image was present in both sets. The number of instances the best image was selected from the WS sets was 36 again, compared to 13 for CS. Moreover, people disagreed more with the images in the CS sets: 2.9 images were deleted from these sets on average compared to 1.6 from the WS sets. There was no apparent trend that people simply preferred the larger result set: 26 votes for the larger set, 22 for the smaller. Hence, the general consensus is that the selection of images obtained in WS is better.

2. *User perception of task performance*: After each task the users were asked if they thought they had succeeded in their performance of the task and also rate potential problems that might have affected their performance. Table 6

Table 5 User effort indicators per task and system

	Task D	Task E	Task F	$\overline{\text{CS}}$	$\overline{\text{WS}}$
Time	10'58"	16'22"	11'56"	12'40"	13'35"
#Images	11.0	18.3	15.7	13.7	16.2
#Queries	10.7	20.3	16.4	16.4	15.3
Manual	8.0	14.0	11.7	11.9	10.6
RF	2.7	6.4	4.8	4.6	4.7

Example illustration task (Task E):
Imagine you are the web designer for an online travel agency called PerfectHoliday. In order to gain more customers, they have decided to hold a competition entitled "Win your dream holiday". They have provided you with the details of the competition (see below) and have asked you to select some images to illustrate the text.
Your task is to nd one main and two additional images that you would place on the webpage along with the competition details. The images should draw people's attention and spark their imagination.

> ***Win your dream holiday!***
> *What if you could make your dream holiday become reality? Where would you go and what would you do? PerfectHoliday is giving you the chance to win that dream! We will be giving away £2000 to the lucky winner for the holiday of their dreams! What would you do with the money? Swim with the dolphins? Stay on a French castle or sail the Mediterranean on a luxurious sailboat? Do you imagine yourself white water rafting in the Alps? Or would a secluded beach with pearly white sands be for you? No matter what your dream holiday looks like, we will make your dreams come true.*
> *To enter this competition, simply send us a description of the perfect holiday before midnight on [...] So don't hesitate! Send your details to [...] and you could be packing your bags!*

Abstract search (Task F):
Imagine you want to take part in a photo competition, where you could win £100 for a picture that depicts the following theme: Dynamic [//Cute]
In order to get ideas for the competition, you want to look for already existing photographs conveying the same theme. Your task is to select at least one images that represents the theme well.

Fig. 7 Task descriptions for Experiment 2

Table 6 User perception of task performance per task and system (performance: higher = better, problems: lower = more problematic)

	Task D	Task E	Task F	$\overline{CS}$	$\overline{WS}$
Performance success	4.4	4.1	4.3	4.2	**4.4**
Did not understand task	4.9	4.9	4.8	5.0	**4.8**
Images not in collection	4.3	3.5	3.6	3.9	3.8
No relevant images returned	4.2	3.6	4.4	4.0	4.1
Not enough time	4.8	4.3	4.8	4.7	**4.5**
Unsure of next action	4.3	4.3	4.2	**4.2**	4.4

reflects the general perception of performance success for each task. The table also highlights the problems that affected the performance (there were no significant differences). The biggest problem encountered was that people thought the images they were looking for were not contained in the collection, followed by the system not returning relevant images. People were slightly less satisfied with their performance for Task E, mainly because they could not find the images they were visualising (i.e. because the images were not in the collection or the system did not return relevant images). Also, time was more of an issue in this task. These results can be expected, since this is the most creative of the three tasks.

Performing a task on WS was generally more successful. The understanding of the task and time were the two issues that had a larger impact on task performance on WS than CS[3]. On the other hand, people's performance was hindered more by an uncertainty of what action to take next on CS. Together with the user comments presented below this indicates that – though a simple concept in principle – providing relevance feedback brings uncertainty as to which images to select for feedback in order to achieve better results. This corroborates similar results in textual information retrieval [16]. On a side note, the implicit negative feedback strategy in WS did not seem to leave people feel out of control. Although negative items could not be reset for groups, that did not have an impact, since the negative feedback was not used for changing the retrieval parameters. For long-term usage of the systems, this would probably become more of an issue.

[3] In Experiment 1, people also tended to agree more with the statement that they had enough time to complete their task on CS: $\overline{CS} = 4.6$, and $\overline{WS} = 4.3$

2. *User satisfaction*: In this section, we discuss the results to the responses concerning user satisfaction with the system in general and the interface features in particular.
   1. *Tasks, search process and retrieved images*: The trend on the user's perception of the tasks themselves is reversed in Experiment 2: the tasks were considered slightly more *clear*, *easy*, *simple*, and *familiar* on WS. As in Experiment 1 there were no significant differences concerning the tasks, however. The search process was once again perceived significantly more *interesting* on WS and the set of images received through the searches were more *complete*. The results for this part are shown in Table 7.
   2. *System and interaction*: There is a clear trend that the participants were more satisfied with WS. They regarded WS to be significantly more *flexible* and the scores for the remaining differentials – *wonderful*, *satisfying*, *stimulating*, *efficient*, and *novel* – were higher for WS as well. CS, on the other hand, was only thought to be *easier*. Table 8 shows the results for these differentials.

      A similar trend is apparent concerning the interaction with the system. People felt more *comfortable* and *confident* while using WS. However, WS was deemed slightly more difficult to *learn to use* but equally easy to *use*.
   3. *Interface support*: In Experiment 2, people were asked how effective they found the interface and rated the contributing features. Table 9 summarises these results. Overall, WS was regarded significantly more *effective*. The three top rated features on WS were that it helped to *organise images*, *explore the collec-*

Table 7 Semantic differential results for the task, search process, and images parts

Differential	$\overline{CS}$	$\widetilde{CS}$	$\overline{WS}$	$\widetilde{WS}$	*P*
Task					
Clear	4.6	5	4.7	5	–
Easy	3.8	4	3.9	4	–
Simple	3.6	4	3.8	4	–
Familiar	3.4	4	3.5	4	–
Search					
Relaxing	3.7	4	3.7	4	–
Interesting	3.6	3	**4.3**	4	0.009
Restful	3.7	4	3.5	3	–
Images					
Relevant	4.0	4	4.1	4	–
Appropriate	4.1	4	4.1	4	–
Complete	3.5	3	**3.8**	4	–

Table 8 Results for the system and interaction differentials, and Likert-scales in the system part

	$\overline{\text{CS}}$	$\widetilde{\text{CS}}$	$\overline{\text{WS}}$	$\widetilde{\text{WS}}$	P
System diffs					
Wonderful	3.3	3	**4.1**	4	–
Satisfying	3.2	3	**4.0**	4	–
Stimulating	3.5	3	**4.3**	4	–
Easy	**4.0**	4	3.8	4	–
Flexible	2.9	3	**4.2**	4	0.004
Efficient	3.3	3	**3.9**	4	–
Novel	3.7	4	**4.4**	5	–
Inter					
In control	3.6	4	3.6	4	-
Comfortable	3.7	4	**4.3**	5	–
Confident	3.1	3	**3.8**	4	–
Likert					
Learn to use	4.1	4	3.9	4	–
Use	3.9	4	3.9	4	–

Table 9 Interface effectiveness

Statement	$\overline{\text{CS}}$	$\widetilde{\text{CS}}$	$\overline{\text{WS}}$	$\widetilde{\text{WS}}$	P
Effective	3.7	4	**4.4**	5	0.032
Analyse task	2.8	3	**4.3**	5	0.001
Explore collection	3.5	4	**4.6**	5	0.001
Find relevant images	4.2	4	4.2	4	–
organise images	2.7	3	**4.7**	5	0.001
Detect/express task aspects	3.0	3	**4.2**	4	0.003

tion, and *analyse the task*. The ordering of features on CS was: *find relevant images*, *explore the collection*, and *detect/express different task aspects*. While both systems were considered equally effective to *find relevant images*, all other features were rated significantly higher on WS.

Table 10 compares the adaptive querying mechanisms in both interfaces: the relevance assessment in CS and the grouping in WS. It turns out that the grouping was considered significantly more *effective* and *useful*. Also note that the relevance assessment was even considered more *difficult* than the grouping.

In open-ended questions the participants were asked to state the most and least useful tools of the interface. The most useful tools in CS were stated as, in order of frequency of responses: textual query (ten responses[4]), QBE search (9), and relevance feedback facility (7). The least useful tools were: result filters for various features (5)[5], relevance feedback (4), and lack of storing facility/overview of selected images (4). It emerged that the relevance assessment was mainly useful for specific searches, for which users stated it as helping them to improve and/or narrow down their search. However, for other tasks people often felt the system returned unexpected results and they were unsure of which items to select to improve the results. In this case, people had to resort to manual search facilities.

In WS, people unanimously liked the grouping facility on the workspace. The three most useful tools in WS included the grouping of images (14), group recommendations (10) and textual queries (5), and the least useful tools were: QBE (4), top 10 window of recommendations (3) and text search (2). This shows that using groups and recommendations was considered more useful than the manual search facilities. Especially the QBE facility was superfluous in this system. There was a plethora of comments about the workspace demonstrating its advantages, such as *grouping was useful to keep track of associated images*, *emphasis was on sorting rather than searching; workspace and groups were used to categorise images and explore those categories further*. The grouping's only disadvantage that became apparent was that it was difficult to remove images from existing groups.

These results support our view that WS, with its grouping and recommendation facility, assists the user in the query formulation process, while removing the need to manually reformulate queries. The picture in CS is quite different: people were divided on the usefulness of the relevance assessments and some still relied heavily on the manual query facilities. On average, people selected 2.4, 3.2, and 3.8 images per relevance feedback iteration for Tasks D, E, and F, respectively. Compared to that, the groups in WS contained 4.9, 4.6, and 4.4 images.

Table 10 Relevance assessment on CS versus grouping on WS

Differential	$\overline{\text{CS}}$	$\overline{\text{WS}}$	P
Easy	3.8	**4.4**	–
Effective	3.3	**4.3**	0.019
Useful	3.7	**4.4**	0.017

[4] This question was asked after each task, thus 24 responses are possible per system.

[5] Apart from the overall results based on both features, the user could look at the individual results for the visual and textual features, respectively.

Table 11 Organisation and information need development results

	Task D		Task E		Task F		AVG		*P*
	CS	WS	CS	WS	CS	WS	CS	WS	
# Selected images	9.6	12.3	17.9	18.6	13.6	17.8	13.7	**16.2**	
Initial idea	3.9	4.6	4.1	4.5	4.0	3.4	4.0	4.2	–
Detect more aspects	2.9	3.6	3.0	3.9	2.8	4.3	2.9	**3.9**	0.05
Satisfied with results	3.3	4.5	3.4	4.1	3.9	4.3	3.5	**4.3**	0.04

So the manual selection process was less productive than collecting the images in groups. Moreover, the grouping process has the additional benefit of supporting a diversifying search by allowing to declare and pursue various task aspects simultaneously.

4. *System rankings*: After completing all four search tasks, the users were again asked to state their preferences for the two systems. 67% liked WS best and the majority also thought it was more effective (42% compared to 8% for CS). CS was clearly easier to learn to use (75%), whereas the ranking for using the systems was balanced (42% for both).

3. *Organisation analysis and information need development*: Experiment 1 has already pointed to differences between users' organisation behaviour depending on the nature of task they performed. Following on the investigation into this dependency, three more types of tasks were introduced in Experiment 2. We observed once more that the more open or complex a task is, the more groups were created on the workspace (1.5, 2.9, and 2.6 for Tasks D, E, and F, respectively). For these types of tasks the organisation was deemed most useful and recommendations were requested more often. Again, the overall results are presented for both experiments below in Sect. 7.2.

   We also analysed the nature of the groups created by the participants for any given task (the detailed results have been omitted due to space limitations). The groups different users created often overlapped in the overall themes, but not necessarily in the images themselves. This shows that groups are definitely task-dependent and hence people would benefit from using and working with other people's groups.

   Moreover, we were interested in the systems' support in developing and fulfilling the user's information need. While their initial idea was clearer on WS, especially for Tasks D and B, they also discovered significantly more task aspects during the search on WS (see Table 11). As could be seen in the performance analysis in Experiment 1 (Sect. 5.1), the category search tasks were more successful on CS. For the tasks in Experiment 2, on the other hand, the participants managed to find a larger selection of images on WS than on CS. The participants were significantly more satisfied with their results across all three tasks, and as we have seen earlier in Table 6 also perceived their overall task performance more successful.

6.3 Summary

All in all, Experiment 2 has essentially reinforced the findings of Experiment 1 regarding the strengths and weaknesses of WS. In addition, it could be shown that the effectiveness for realistic, design-based tasks is actually better on WS. There was also more evidence that the grouping and recommendations caused less confusion and were more natural to the users than the relevance feedback approach. Before drawing the final conclusions, we will provide the combined results of Experiments 1 and 2 in the next section.

7 Combined results

In this section, we present the combined experimental results with an emphasis on a task-based comparison. It provides a discussion on users' perception of task characteristics and performance in order to analyse the specifics of each task. Further, a summary of the organisation analysis should help to clarify how people useed the workspace for all tasks performed on WS. Last but not least, we present the overall results concerning usability and user satisfaction for all 24 users.

7.1 Task analysis

We have created a variety of realistic tasks, ranging from category search, an image-based theme search, abstract topic search, illustration task and an open design task. The tasks were designed to vary in terms of complexity, degree of abstraction and creativity. The participants

Table 12 Summary of task descriptions and number of user samples per system

Task	Description	Objective	CS	WS
A	Simple or focused category search tasks	Find as many images as possible for the specified topic	6	6
B	Complex or multifaceted category search tasks	Find as many images as possible for the specified topic	6	6
C	Design task	Choose 3–5 images to design a leaflet	0	12
D	Theme search tasks	Choose one image to complement a provided set of three images of a specific theme	8	8
E	Illustration task	Choose three images to illustrated a provided piece of text or advertising slogan	8	8
F	Abstract topic search tasks	Choose one image of a specified abstract topic	8	8

Table 13 Semantic differentials about task perception per task

Differential	Task A	Task B	Task C	Task D	Task E	Task F	$\overline{CS}$	$\overline{WS}$	*P*
Task									
Clear	4.8	4.8	4.8	4.8	4.8	**4.4**	4.7	4.7	–
Easy	**4.4**	**4.3**	3.9	3.9	3.8	3.7	4.0	4.0	–
Simple	**4.7**	**4.7**	3.9	3.6	3.6	3.7	4.0	4.0	–
Familiar	3.6	3.9	3.8	3.5	3.4	3.4	3.6	3.6	–
Search									
Relaxing	4.0	**4.5**	3.9	3.6	3.8	3.8	4.0	3.8	–
Interesting	3.9	3.9	**4.1**	3.8	**4.3**	3.9	3.6	4.3	0.006
Restful	3.3	3.4	3.6	3.6	3.6	3.7	3.6	3.5	–
Images									
Relevant	**4.4**	3.9	4.3	4.2	4.0	4.1	4.1	4.1	–
Appropriate	**4.5**	4.0	3.9	4.1	4.0	4.2	4.1	4.2	–
Complete	3.5	3.6	3.3	3.8	3.6	3.8	3.4	3.9	0.006

confirmed that they were familiar with these types of tasks and that they encountered similar tasks in their own time. The tasks are described in Sects. 4.3 and 2, and are summarised in Table 12. The number of users per task is also specified there.

Through the analysis of task characteristics and the resulting performance, we hope to identify the types of tasks that each system is most appropriate for. So first, we look at the task characteristics from the users' perspective. Table 13 (scores from 1 to 5, higher = better) shows that the tasks were perceived equally clear with an exception of the abstract search task. The category search was considered the easiest and simplest, followed by the design task, the theme search and the illustration task on a par, and finally the abstract search. This shows that task complexity is related to the decision making process required. The decision making process was less crucial in the category search tasks since the objective was to find as many images as possible from a (well-defined) given category. Furthermore, the search process was considered the more interesting, the more creativity was asked for in a task. However, people's expectation of the appropriateness of the retrieved images was also higher for the creative tasks. Thus, the more specific the task, the more people thought the system helped to retrieve the right images (relevant and appropriate).

It also emerged that the perception of task complexity sometimes varied depending on which system the task was performed on. Most importantly, the search process was considered significantly more *interesting* on WS for all tasks. The most notable difference was between the two category search tasks. Tasks A and B were considered more *simple* on CS than WS: 4.8 and 4.5, respectively. Task A led to the most *stressful* search process on WS (CS: 4.5, WS: 3.5) and the images were considered less *relevant* (CS: 4.7, WS: 4.2) and *appropriate* (CS: 4.7, WS: 4.3). The opposite was true for the complex categories (3.7 for both image differentials on CS; 4.2 and 4.3 for WS, respectively)[6]. In addition, the judgement of Task D changed depending on which system was used. It was considered more *clear* (CS: 4.6, WS: 4.9) and *easy* (CS: 3.6, WS:4.3) and less *complex* (CS:3.6, WS:4.0) when using WS.

Next, the task performance has been looked at in the results analysis sections for Experiments 1 and 2 (Sects. 5 and 6.2, respectively). We briefly reiterate our

[6] It is also interesting to note that these two differentials scored the same on average for both the simple and complex categories on WS.

observations on the users' perception of their success in performing a given task. People were least happy with their performance in the more creative tasks, mainly due to not having had enough time to complete the task. People were more satisfied with their performance on WS, although time was a bigger issue here. On the other hand, uncertainty about the next action affected their performance more on CS. The actual task performance for the category search tasks was consistently better on CS. However, there was no correlation between actual and perceived task performance.

Finally, we analysed the amount of user effort required to solve a task (for Experiment 2). Most time was spent on the illustration task, reconfirming user's perception on task performance in this respect. However, more images were selected and more queries were issued during the course of this task. Adaptive queries in the form of relevance feedback iterations or group recommendations were considered especially valuable for this type of task. WS helped the user to select more images for all three Experiment 2 tasks.

To conclude, we could see differences in the perception of tasks and the actual effort required both depending on the nature of the task as well as the system being used. In summary, CS seems to be good for quickly finding many images for a specific/narrow topic. The strengths of WS show particularly for more complex or creative tasks. Especially if the information need is vague in the beginning, the grouping facility on WS allows the user to explore the collection and to discover and express different task aspects. Therefore, users of WS are encouraged to diversify their search. The workspace makes it possible to make a more informed decision on the final images selected from a larger set of alternatives.

7.2 Organisation analysis

In this section, we summarise people's organisation and the nature of their information need for all tasks performed on WS. Table 14 shows the relevant data per task. We can observe that the number of groups created corresponds to the number of facets the users detected and followed up on. From this perspective, Tasks A and D were represented by a single facet (approximately), while the other tasks had about 3–4 facets. Tasks C–F had clear instructions on how many images had to be selected. These targets are closely reflected in the number of groups created, with the exception of Task F. The target for Task F was to select only one image, but was represented by 2.6 groups on average. Since the topic for Task F was abstract (especially compared to Tasks A and D), people explored several alternatives, which correspond to the number of groups they created. Task B is also interesting in this respect. Although the target was the same as in Task A, namely to find as many images as possible, people created more groups for the topics in Task B, which were more complex than the topics in Task A. Hence, the number of facets is influenced by two factors: (1) the complexity of the task and (2) the number of images that were required for the task.

The nature of the underlying information needs is captured by asking how clear people's initial idea (before starting the search) was and if they detected more aspects while searching. The responses for their initial idea are again an indicator for how focused the tasks were perceived. Tasks F and C have the lowest score of initial idea, and are indeed more open to interpretation than the other tasks. As mentioned before, Task F is the most abstract and Task C the most creative. Interestingly, there is a relationship between the scores of initial idea and task aspects: they are roughly inversely proportional. So, the less defined their initial idea, the more aspects users detected during the search and vice versa. Task B is the only exception: the information need was well-defined but people also detected more aspects. This is not too surprising, because people can think of many images for a category such as "African Wildlife" from the top of their head – unlike the abstract topic of Task F. Since these topics comprise a large number of facets (at least 4.3 that were detected on average) people can still find some more during the search they had not thought of before.

The large difference in result satisfaction between Tasks A–C and Tasks D–F can possibly be explained by the improved retrieval system in Experiment 2. Still, we can see that the creative tasks (Task C in Experiment 1 and Task E in Experiment 2) have the lowest scores compared to the other tasks in the same experiment. We believe that this is due to higher expectations for these tasks. People are instructed to create a composition of images rather than select images with a specified requirement. As seen above, time restrictions were an issue affecting their performance satisfaction, probably affecting their satisfaction with the results as well.

Finally, the organisation feature was regarded as very useful. The only exception was for finding a large number of images from a focused topic. In fact, CS was generally preferred for this task.

7.3 User satisfaction with systems

In both Experiments 1 and 2, the participants were asked to rate the system they had just used in the post-search questionnaires. These results are given per experiment above. Nonetheless, we provide the combined results for

Table 14 Organisation and information need development for all tasks on WS

	Task A	Task B	Task C	Task D	Task E	Task F
# Groups	1.2	4.3	4.4	1.5	2.9	2.6
Initial idea	4.5	4.3	**3.7**	4.6	4.5	3.4
Detected more aspects	**3.0**	4.7	4.3	3.6	3.9	4.3
Satisfied with results	3.7	3.5	**3.0**	4.5	**4.1**	4.3
Organisation useful	**3.0**	4.8	4.4	4.6	4.8	4.7

all 24 users of both experiments in this section, since a larger sample size leads to more reliable results.

Table 15 shows the results for the system part in the post-search questionnaires. The participants considered CS *easier* than WS, while they considered WS to be significantly more *stimulating*, *flexible*, and *novel*. The scores for the remaining differentials, *wonderful*, *satisfying* and *efficient*, were generally higher for WS as well. While using the system, people felt more *comfortable* and *confident*. However, WS was deemed more difficult to *learn to use* and to *use*. Table 16 shows the users' preferences of systems for the statements asked in the exit questionnaire. 67% liked WS best and the majority also thought it was more effective. CS was clearly easier to learn to use, whereas the ranking for using the systems was relatively balanced.

In open-ended questions, the participants were asked for their opinion on what they liked or disliked about each system. The advantages listed for CS were that it was easy to use, fast and efficient especially for specific searches: e.g. *easy to drill down and find 1 or 2 images you were looking for at the start*. Its disadvantages included that the users felt they did not have enough control over the search and that its interface was less intuitive (*too abstract*, *slightly confusing*). People appreciated WS as an organising tool. The workspace enabled them to plan their tasks and pursue alternative search threads, without losing the overview of intermediate results and searches: e.g. *ability to group and then follow alternative search threads*, *very useful if looking for a variety of different images in the same topic*. In addition, the system's flexibility and more control options were noted as advantages, e.g. *it allowed flexibility* [...] *therefore I selected more, then dispensed with those that weren't useful*. In Experiment 1, 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 were consequently improved for Experiment 2. The recommendation quality was improved by taking textual annotations into account. The handling of the groups and images within groups was changed so that the system now automatically arranges the layout of the images in a group. Consequently, none of these issues resurfaced in Experiment 2.

Table 15 Results for the system part (Experiment 1 + 2)

Differential	$\overline{\text{CS}}$	$\widetilde{\text{CS}}$	$\overline{\text{WS}}$	$\widetilde{\text{WS}}$	*P*
Wonderful	3.4	4	**4.1**	4	–
Satisfying	3.5	4	**4.0**	4	–
Stimulating	3.4	3	**4.3**	4	0.007
Easy	**4.3**	4	3.8	4	–
Flexible	2.9	3	**4.1**	4	0.000
Efficient	3.6	4	**3.9**	4	–
Novel	3.4	3	**4.3**	5	0.005
In control	3.8	4	3.8	4	–
Comfortable	4.0	4	**4.4**	5	–
Confident	3.6	4	**4.0**	4	–
Learn to use	**4.3**	4	4.1	4	–
Use	**4.2**	4	3.9	4	–

Table 16 Comparison of system rankings

System	Learn	Use	Effective	Liked best
CS	14 (58%)	10 (42%)	5 (21%)	4 (17%)
WS	5 (21%)	11 (46%)	11 (46%)	16 (67%)
Same	5 (21%)	3 (13%)	8 (33%)	4 (17%)

8 Summary

Although a workspace has been introduced in a few information retrieval systems before albeit with limited functionality, for instance ImageGrouper [17] and SketchTrieve [18], its usefulness has not been evaluated formally yet. Above all, it has not been studied how results organisation assists image retrieval. With this two-stage experiment, involving 24 participants on a variety of real-life search tasks, we aimed at filling this gap and answer the following questions: How was the workspace used? What influence did the task have on this? Does it help to conceptualise tasks? Does it help overcome the query formulation problem? These are the answers we found for our specialised domain of results organisation for image retrieval:

How was the workspace used and what influence did the task have on this? As determined in the organisation and information need analysis, the workspace was used to create different groupings that reflected differ-

ent semantic facets of the task. These facets were often overlapping amongst the users for the same task.

In addition, we found a correlation between task characteristics and organisation behaviour. The workspace is most useful for exploratory searches with vague information needs or complex, multi-faceted tasks. Possible explanations include that it helps to analyse the task better, discover more aspects of the task than initially anticipated, and explore the collection better, which was indicated in the questionnaires.

On the other hand, CS was better for tasks that required selection of a large number of images for a very specific topic. However, the organisation was still deemed useful for focused tasks, because it helped to maintain a better overview and hence better comparison opportunities of the selected images. For these tasks, the focus shifts naturally to selecting images with good quality rather than the pure quantity of images. In the users' eye this was a more realistic goal of image searching tasks.

Does it help to conceptualise tasks? Grouping search results on the workspace incites the user to organise results for their search/work task. This enabled the users to break up their overall search task into a small set of individual search tasks. Hence, the grouping process has the benefit of allowing the user to explore the task. People can pursue a progressive search strategy by following multiple search threads simultaneously, while maintaining a constant overview of intermediate results and searches. The groups are equivalent to task aspects, and the search threads are equivalent to trains of thought. This shows that the workspace *helps users conceptualise and diversify their tasks better* (experimental hypothesis 2).

Does it help overcome the query formulation problem? The grouping facility was not only considered easier, more effective and useful than the relevance feedback approach in CS, but was praised unanimously in open-ended questions. In addition, the relevance feedback facility caused more confusion. It became apparent that providing relevance feedback brings uncertainty as to which images to select for feedback in order to improve the results. Hence, people relied more on the manual query facilities on CS than WS. Although both systems have the same underlying retrieval mechanism, the workspace approach is more successful at eliciting constructive feedback while hiding the internals of the retrieval mechanism. Since the groups are equivalent to task aspects, users find it easier to categorise images into these aspects and interpret the system's results accordingly. Consequently, people selected more images for feedback and requested more recommendations on WS than RF iterations on CS. Thus, one can conclude that the grouping process is *better at overcoming the query formulation problem* (hypothesis 3).

The ability to group search results together with the recommendation facility, has increased the effectiveness of the system. The required effort to complete a task was lower on WS: less queries were issued to find a larger selection of images. In particular, users created less manual queries but issued more system recommendations. The participants also perceived their performance as more successful on WS and the interface was perceived significantly more effective for completing the tasks. This shows that the workspace *increased the effectiveness of the search* (hypothesis 1).

These observations have led us to accept all three experimental hypotheses. However, this study also helped to identify the limitations of the workspace. WS was more difficult to use and the cognitive effort required to solve a task was higher. This was reflected in the questionnaire responses; in particular users had more difficulty in understanding the task and it took longer to complete it. However, the longer learning period and increased cognitive effort is not perceived as a disadvantage of WS; after all, 16 people preferred WS over CS. More importantly, we found evidence that attributed the prolonged search session to the system's ability to support the user in exploring the tasks from different perspectives. As mentioned before, people were able to diversify their search better and follow up on multiple trains of thought simultaneously. Still, one has to keep in mind that it takes longer to become familiarised with this interface, although we strived to make its operation as intuitive as possible by using standard commands which the user may already be familiar with wherever possible.

Finally, we did not explore the use of WS for collaborative image retrieval. On the workspace, people leave footprints of their activities behind for later usage. We also observed that people's groups overlapped in their overall themes, which could be exploited in a collaborative context. Such a feature will be explored in future studies.

9 Conclusion

In this paper, we have established the usefulness of the workspace system for image retrieval. We have created a realistic experimental study, in which design professionals performed a variety of realistic search tasks. Based on the results of this experiment, we argue that the workspace is an indispensable tool in an image retrieval system. It is used for organising the results according to the different aspects or facets of the task. This helps

users greatly in analysing and exploring the task as well as the collection. Moreover, the workspace supports a more intuitive search process and helps to overcome the query formulation problem. All these factors lead to a more effective and enjoyable search experience.

Acknowledgements We would like to thank all participants of this study for their time and valuable feedback. Many thanks also to our anonymous reviewers for useful suggestions. This work was partially supported by the EPSRC (Grant ref: EP/C004108/1). This publication only reflects the authors' views.

References

1. Urban, J., Jose, J.M., van Rijsbergen, C.J.: An adaptive technique for content-based image retrieval. Multimed. Tools Appl. (2006) (in press)
2. Urban, J., Jose, J.M.: EGO: A personalised multimedia management and retrieval tool. Int. J. Intell. Syst. (IJIS), Special Issue on 'Intelligent Multimedia Retrieval', vol. 21, no. 7, pp. 725–745 (2006)
3. Garber, S.R., Grunes, M.B.: The art of search: a study of art directors. In: Proceedings of the ACM International Conference on Human Factors in Computing Systems (CHI'92), pp. 157–163 (1992)
4. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices in the digital newspaper photo archive. Information Retrieval, vol. 1, no. 4, pp. 259–285 (2000)
5. The Benchathlon Network, Home of CBIR Benchmarking. Available online at http://www.benchathlon.net/
6. The CLEF Cross Language Image Retrieval Track (ImageCLEF). Available online at http://ir.shef.ac.uk/imageclef/
7. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth, London (1979)
8. Jose, J.M., Furner, J., Harper, D.J.: Spatial querying for image retrieval: a user-oriented evaluation. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 232–240. ACM Press, New York, (1998)
9. Urban, J., Jose, J.M.: Evidence combination for multi-point query learning in content-based image retrieval. In: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04), pp. 583–586 (2004)
10. Rui, Y., Huang, T.S.: Optimizing learning in image retrieval. In: IEEE Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 236–245. IEEE Computer Society Press (2000)
11. Miller, G.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 81–97 (1956)
12. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham, London (1992)
13. Berkley's list of Corel CD names, images keywords and captions, Berkley's Digital Library Project. Available online at http://elib.cs.berkeley.edu/photos/corel/
14. Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill, Tokyo (1983)
15. Zhou, X.S., Huang, T.: Relevance feedback in image retrieval: a comprehensive review. ACM Multimed. Syst. J., Special Issue on **8**(6), 536–544 (2003)
16. Beaulieu, M., Jones, S.: Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. Interact. Comput. **10**(3), 237–248 (1998)
17. Nakazato, M., Manola, L., Huang, T.S.: ImageGrouper: a group-oriented user interface for content-based image retrieval and digital image arrangement. J. Vis. Lang. Comput. **14**, 363–386 (2003)
18. Hendry, D.G., Harper, D.J.: An informal information-seeking environment. J. Am. Soc. Inf. Sci. **48**(11), 1036–1048 (1997)