Multimed Tools Appl (2013) 64:695–716
DOI 10.1007/s11042-011-0951-6

The Mosaic Test: measuring the effectiveness of colour-based image retrieval

William Plant · Joanna Lumsden · Ian T. Nabney

Published online: 4 January 2012

Abstract A variety of content-based image retrieval systems exist which enable users to perform image retrieval based on colour content—i.e., colour-based image retrieval. For the production of media for use in television and film, colour-based image retrieval is useful for retrieving specifically coloured animations, graphics or videos from large databases (by comparing user queries to the colour content of extracted key frames). It is also useful to graphic artists creating realistic computer-generated imagery (CGI). Unfortunately, current methods for evaluating colour-based image retrieval systems have 2 major drawbacks. Firstly, the *relevance* of images retrieved during the task cannot be measured reliably. Secondly, existing methods do not account for the creative design activity known as *reflection-in-action*. Consequently, the development and application of novel and potentially more effective colour-based image retrieval approaches, better supporting the large number of users creating media for use in television and film productions, is not possible as their efficacy cannot be reliably measured and compared to existing technologies. As a solution to the problem, this paper introduces the *Mosaic Test*. The Mosaic Test is a user-based evaluation approach in which participants complete an image mosaic of a predetermined target image, using the colour-based image retrieval system that is being evaluated. In this paper, we introduce the Mosaic Test and report on a user evaluation. The findings of the study reveal that the Mosaic Test

W. Plant (✉) · J. Lumsden · I. T. Nabney
Computer Science, School of Engineering and Applied Science,
Aston University, Birmingham, UK
e-mail: plantwr1@aston.ac.uk

J. Lumsden
e-mail: j.lumsden@aston.ac.uk

I. T. Nabney
e-mail: i.t.nabney@aston.ac.uk

overcomes the 2 major drawbacks associated with existing evaluation methods and does not require expert participants.

Keywords Image retrieval · Image databases · Content-based image retrieval · Query-by-sketch · Query-by-colour · Performance evaluation

1 Introduction

A variety of content-based image retrieval systems—such as QBIC (query by image content) [2] or MARS (Multimedia Analysis and Retrieval System) [11]—and image retrieval systems popular amongst design communities (e.g., Google Images [3], Bing Images [9] and iStockPhoto [7]) enable users to perform image retrieval based on colour content—i.e., colour-based image retrieval. Colour-based image retrieval is an important tool for users retrieving images from a database with specific colour compositions, as it is often difficult to express the colour layout of an image using just a few keywords. For the production of media for use in television and film, colour-based image retrieval is useful for retrieving specifically coloured animations, graphics or videos from large databases, through comparison of user queries with the colour content of (automatically extracted) key frames [21]. It is also useful to graphic artists creating life-like computer-generated imagery (CGI) for films and television.

Unfortunately, existing methods used to evaluate the effectiveness of colour-based image retrieval systems typically have 2 major drawbacks. Firstly, the *relevance* of images retrieved during a user-based task (such as finding all images from a database 'relevant'to a target image) cannot be reliably measured since there is no objective measure of relevance. Whilst existing evaluation approaches adopt manually created relevance assessments (e.g., those listed in a *ground-truth*), such judgements are highly subjective and can vary greatly between 2 or more human assessors. Secondly, highly creative individuals such as those involved in media production, assess the suitability of retrieved videos and images relative to the context of a creative project (for which the retrieved video or image is intended). This type of activity is known as *reflection-in-action* [16]: current evaluation approaches do not support nor assess the ability of a given colour-based image retrieval system to support reflection-in-action. Consequently, no method currently exists for reliably and *meaningfully* evaluating the appropriateness of colour-based image retrieval systems. Therefore, the development and application of novel and potentially more effective colour-based image retrieval approaches, better supporting the large number of users creating media for use in television and film productions, is not possible as their efficacy cannot be reliably measured and compared to existing technologies.

This paper introduces the Mosaic Test which has been developed to address these issues, by providing a reliable method by which to meaningfully evaluate colour-based image retrieval systems. The Mosaic Test is a user-based evaluation system in which participants complete an image mosaic of a predetermined target image using the colour-based image retrieval system under evaluation. The time and workload required for participants to complete this creative task, as well as the city block (or L_1) distance between MPEG-7 colour structure descriptors [17] of the target images and user-generated image mosaics, are used to assess the effectiveness of

the colour-based image retrieval system being tested. In this paper, we additionally report on a user study that was conducted to evaluate the Mosaic Test.

The outline of the paper is as follows. In Section 2 we describe the background of colour-based image retrieval, in particular the need for colour-based image retrieval in media production, how it is implemented and the drawbacks associated with current methods for evaluating the performance of colour-based image retrieval systems. Section 3 details the Mosaic Test and how it is able to overcome the drawbacks of existing colour-based image retrieval evaluation approaches. Section 4 outlines the conducted user study to evaluate the Mosaic Test. Finally, Section 5 presents and discusses the results of the user study, whilst Section 6 concludes the paper.

2 Background

In this section we describe the need for colour-based image retrieval in media production, how it can be implemented and the drawbacks associated with current methods for evaluating the performance of colour-based image retrieval systems. In Sections 2.1 and 2.2, we respectively discuss the need for colour-based image retrieval in media production, and how colour-based image retrieval can be implemented. In Sections 2.3 and 2.4, we describe the only existing method applied to evaluating a colour-based image retrieval system, and other approaches used to evaluate more general image retrieval systems. Finally, we discuss the 2 fundamental drawbacks of the described evaluation approaches in Section 2.5.

2.1 Colour-based image retrieval in media production

Colour-based image retrieval techniques are not limited exclusively to image databases and image retrieval. Colour-based image retrieval can also be applied for retrieving video content, such as animations, graphics or real-world footage, on the basis of colour. Video data can be hours in length, and thus manually inspecting and retrieving specifically coloured video footage or animations from a large video database manually is an impossible task. Consequently, much research has been conducted into content-based techniques for video retrieval. A popular approach commonly investigated by researchers in the field is *video abstraction*, in which the content of a video is summarised using a small set of stationary images known as keyframes [21]. These keyframes can then be indexed by colour-based image retrieval systems, enabling those involved in media production to retrieve videos from a database the basis of colour.

Colour-based image retrieval is also an important tool for graphic artists, often responsible for creating life-like computer-generated imagery (commonly referred to as CGI) for use television and film. Here, graphic artists are required to apply suitably coloured and textured images to objects and characters existing in virtually generated environments in order make them appear more realistic and life-like. We found evidence of this when we examined keyword queries submitted to a popular online texture repository. Users of CG Textures [22], a web site offering images of real world textures for use in CGI, entered keywords such as “tile blue” or “red rust” when searching for requisite images.

2.2 Implementing colour-based image retrieval

It is clear that in order to retrieve images based on their colour content, the most important issue to address is how that colour content should be represented: how precise should the description of colours be, what threshold should be used to determine the presence of a colour, whether spatial distribution of colour is relevant. The other key question is how the similarity between colour representations should be measured.

Colour-based image retrieval systems typically adopt either of the *query-by-colour* or *query-by-sketch* paradigms in order to facilitate image retrieval on the basis of colour. For the query-by-colour paradigm, users formulate queries by selecting requisite colours from a graphical colour palette (and, in some systems, specifying the ratio of his or her selected colours). An example of a graphical colour palette can be seen at the top of the left hand colour-based image retrieval system shown in Fig. 1. In the case of the query-by-sketch paradigm, users are asked to generate a sketched example of the images he or she requires, using a simple drawing tool. An example of a simple drawing tool for query-by-sketch can be seen at the top of the right hand colour-based image retrieval system shown in Fig. 1. For both paradigms, the system extracts a colour descriptor from the queries of users which are compared with the colour descriptors extracted from each database image (which summarises the colour content of that image in a compact form to reduce storage and processing overheads) using some adopted distance metric. The database images with colour descriptors 'closest' (according to the adopted distance metric) to a query are returned by the system and displayed to users so that they may browse the images from the database most relevant to their colour requirements.

The fundamental difference between the colour descriptors used in query-by-colour and query-by-sketch is that those used in the latter must contain information

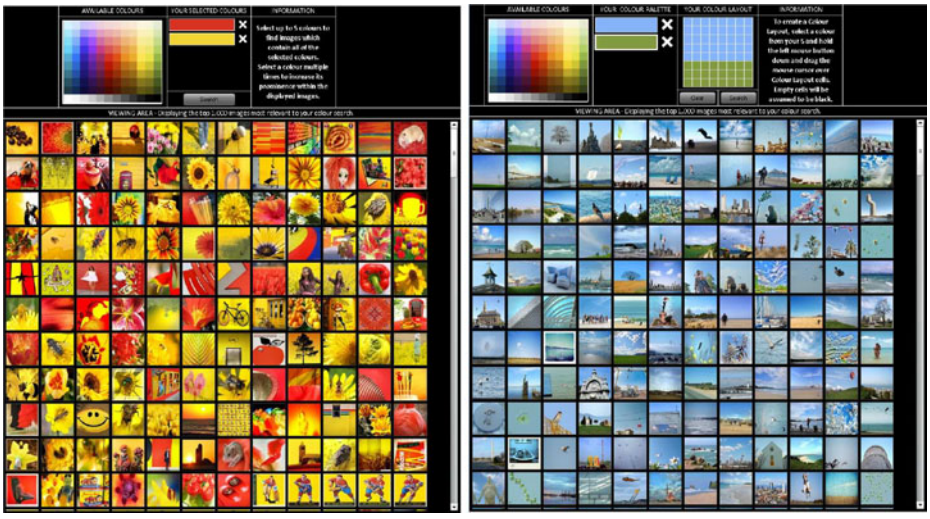

Fig. 1 Screenshots of 2 colour-based image retrieval systems adopting the query-by-colour (*left*) and query-by-sketch (*right*) paradigm

regrading the spatial distribution colour within images (i.e., the contained colours and the location of each within the image) whilst the former does not (i.e., only information regarding the contained colours is required). Whilst a variety of techniques exist for extracting the overall colour distribution in an image (such as the MPEG-7 dominant colour descriptor [17]), the most widely implemented method is via the use of a colour histogram [19]. A colour histogram contains a normalised pixel count for each unique colour in the colour space. To minimise storage and processing requirements, the number of colours in the colour space is typically reduced through the technique of colour quantisation [19]. The colour histogram H for a given image I can be formally defined as $H_I = [B_1, B_2, ..., B_j, ..., B_n]$ where n is the number of distinct colours in the colour space and B_j is the histogram bin containing the number of pixels in image I that are of colour j. These bin values are typically normalised by dividing the corresponding pixel count by the total number of pixels in image I. For colour-based image retrieval systems adopting the query-by-colour paradigm, a colour histogram can be generated from the user query and compared to colour histograms extracted from database images using a variety of distance metrics, including the widely used L_1 and L_2 measures shown in (1) and (2) respectively.

$$L_1(A, B) = \sum_{i=1}^{n} |A_i - B_i| \tag{1}$$

$$L_2(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{2}$$

There are also several colour descriptors that summarise the spatial distribution of colour content within an image [1], making them suitable for use in colour-based image retrieval systems adopting the query-by-sketch paradigm. The auto colour-correlogram (ACC) of an image can be described as a table indexed by colour pairs, where the k-th entry for colour i specifies the probability of finding another pixel of colour i in the image at a distance k. For the MPEG-7 colour structure descriptor (MPEG-7 CST), an 8×8 pixel sliding window moves across the pixels in an image in the HMMD (Hue, Min, Max, Diff) colour space [17] (quantised to 256 colours). This can be seen in Fig. 2a. With each shift of the structuring element, if a pixel with colour i occurs within the block, the total number of occurrences in the image for colour i is incremented to form a colour histogram. The distance between 2 MPEG-7 CSTs or 2 ACCs can be calculated using the L_1 distance metric. Finally, the MPEG-7 colour layout descriptor (MPEG-7 CL) [17] divides an image into 64 blocks (as shown in Fig. 2b), and calculates the dominant colour of the pixels within each block in the YC_bC_r colour space [17]. The cumulative distance between the colours of corresponding blocks forms the measure of similarity between 2 MPEG-7 CL descriptors.

2.3 Evaluating colour-based image retrieval systems

Faloutsos et al. [2] evaluation of the QBIC system is the only example of an attempt to measure the effectiveness of a colour-based image retrieval system in the literature. Faloutsos et al. supplied users with a target image, and asked them to mark each image in a database of approximately 1,000 images as either relevant or

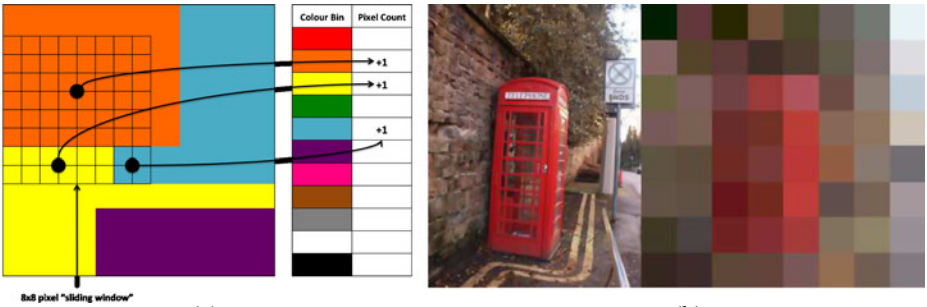

Fig. 2 **a** Describes the 'sliding window' approach of the MPEG-7 CST descriptor. At its current location, each of the 64 (8×8) pixels in the window are one of 3 colours. The counts for these 3 colours in the colour space are therefore incremented by 1. **b** Shows an image of a telephone box alongside its corresponding MPEG-7 colour layout descriptor

irrelevant with respect to the supplied target image. It is unclear whether the users were asked to form this relevance assessment based on colour or semantic content. The users were then asked to submit a single colour query to the QBIC system that they believed would be sufficient to retrieve the target image from the database. Using the top 20 images returned by QBIC as the result of the user's query, Faloutsos et al. calculated the average rank ($AVRR$) of all the relevant images occurring in the result set and the ideal average rank ($IAVRR$). The ideal average rank can be defined as $IAVRR = (0 + 1 + \cdots + (T - 1))/T$, where T is the total number of images in a database relevant to a target image. According to these definitions, an effective colour-based image retrieval system will achieve an $AVRR$ close to the $IAVRR$ value. In the study of Faloutsos et al., these values were averaged over ten different target images.

2.4 Evaluating image retrieval systems

In the previous sub-section, we have highlighted the only existing approach to measuring the effectiveness of a colour-based image retrieval system. In the field of image retrieval, however, there exist a number of techniques which could be adopted for colour-based image retrieval. The colour-based image retrieval evaluation method adopted by Faloutsos et al. is based on the *precision and recall* measures (shown in (3) and (4), respectively), that are commonly reported together when evaluating content-based image retrieval systems. These measure the relevance of the images returned by a content-based image retrieval system. Amongst others, a fundamental problem with the precision and recall measures is that they fail to account for the usability of image retrieval systems.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \tag{3}$$

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}} \tag{4}$$

There are, however, 2 popular evaluation approaches adopted in the image retrieval domain which do require a high degree of user interaction (and thus account for system usability). These are the *category search* and *target search* [12] tasks. For the category search task, participants are shown a target image or keyword and instructed to find all semantically relevant images in a database (using the image retrieval system being tested) within a time limit. The number of images found within the time limit is then used as a measure of the effectiveness of a system for image retrieval. For a target search, users are asked to retrieve a specific image from a database within a time limit (using the content-based image retrieval system to be tested), with the time taken used as a measure of the system's performance. Unfortunately, as we describe in the next sub-section, the approaches described here are unsuitable for evaluating the effectiveness of a colour-based image retrieval system.

2.5 Drawbacks of existing evaluation methods

Each of the existing evaluation methods described in the previous 2 sub-sections incur at least one of 2 major drawbacks. The first issue that can relate to these evaluation methods is the manner in which image relevance is measured. Recall that in the evaluation method of Faloutsos et al., users were required to manually label all images in a database as either relevant or irrelevant to a given target image. Given that the magnitude of image databases has grown dramatically in recent years, such a task would prove extremely time-consuming and arduous for users today.

An alternative to asking users to manually assess image relevance pre-test would be to adopt an existing database, such as [6, 15], which has a corresponding ground truth (i.e. a list of all images in the database relevant to a set of target images). Unfortunately, however, the relevance judgements used in creating such a ground truth are not only highly subjective, but also typically based on high-level content (i.e. what is actually in the image) rather than much lower-level colour relevance (or similarity). These existing image databases and their corresponding ground truths are therefore inappropriate for evaluating the effectiveness of colour-based image retrieval systems. The inability to reliably assess the relevance of images retrieved using a colour-based image retrieval system is the first of 2 major drawbacks associated with existing evaluation methods.

The second fundamental drawback associated with the existing evaluation methods, is that they all fail to reflect accurately the manner in which creative users assess the suitability of images retrieved from an image database for use in a project—that is, they fail to reflect the importance of reflection-in-action [16]. As a result of the 2 drawbacks described above, no method currently exists for reliably evaluating colour-based image retrieval systems. In the next section we shall propose a new evaluation method that does address both of these issues.

3 The Mosaic Test

The Mosaic Test has been developed to address the problems described in the previous section—that is, to provide a reliable means by which to meaningfully evaluate colour-based image retrieval systems. The Mosaic Test is a user-based

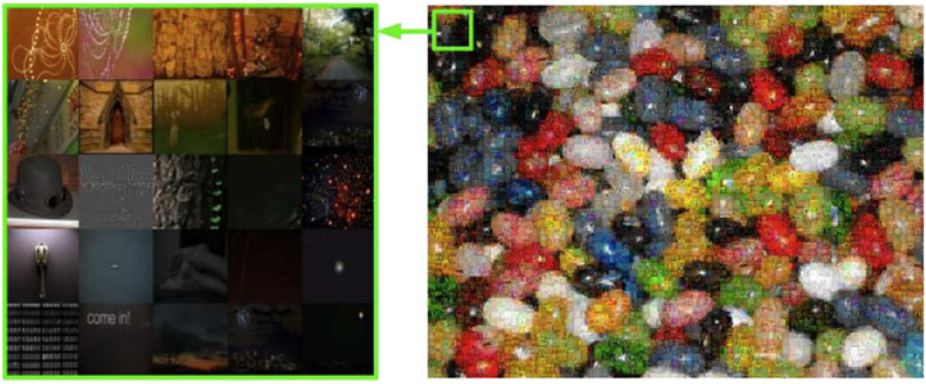

Fig. 3 An example of an image mosaic. The region *highlighted green* in the image mosaic (*right*) has been created using the images shown (*left*)

evaluation in which participants complete an image mosaic of a predetermined target image using the colour-based image retrieval system that is being tested. The Mosaic Test is supported by a Mosaic Test Tool which enables participants of a Mosaic Test to produce an image mosaic manually by selecting images from a database; we describe this tool here. We also describe how colour-based image retrieval effectiveness data— namely, the time, self-reported workload required to complete the image mosaic, and relevance of the images used in an image mosaic—is extracted by the Mosaic Test Tool to conduct a Mosaic Test.

3.1 Image mosaics

An image mosaic is a form of computer-generated art [18]. A target image is divided into cells, each of which is replaced by a small image with similar colour content. Viewed from a distance, the smaller images appear collectively to form the target image; viewing an image mosaic up close reveals the detail contained within each of the smaller images. Image mosaics were first devised by Silvers [18], who describes a system for automatically creating an image mosaic. More recent examples of systems using the colour content of target and database images to automatically create image mosaics are presented in [24] and [10]. An example of an image mosaic can be seen in Fig. 3.

3.2 Measuring effectiveness from manually created image mosaics

The Mosaic Test, which requires participants to create an image mosaic manually, is able to overcome the 2 fundamental drawbacks of existing methods that evaluate the effectiveness of colour-based image retrieval systems. We define *colour-based image retrieval effectiveness* as a measure of the time and workload required by users of a given system to retrieve images (or indeed videos via key frame retrieval) that closely match their colour requirements from a database.

Supported by the Mosaic Test Tool, the Mosaic Test measures each component of this effectiveness metric. As a measure of time, the Mosaic Test uses the number of seconds required by a participant to complete their image mosaic. To account for

the workload element of our definition, each participant is asked to complete the NASA-TLX (task load index) [4] immediately after testing the colour-based image retrieval system: this returns an overall workload score (based on the mean ratings of 6 scales: effort, mental demand, temporal demand, physical demand, frustration and performance) which provides a subjective measure of the workload experienced by users whilst creating an image mosaic with the system under evaluation.

The relevance of the retrieved images used in an image mosaic can be measured reliably by comparing the user-generated image mosaic and the target image. This automatic calculation of image mosaic relevance is explained further in Section 4.1. Participants are able to perform reflection-in-action [16] by adding an image to their image mosaic to assess its suitability (and removing it afterwards, if necessary).

3.3 Mosaic Test Tool

As described in Section 3.1, image mosaics are typically created automatically by a computer program that analyses and compares the colour content of a cell in the target image and images in a database. For the Mosaic Test, however, participants are asked to manually create an image mosaic of a predetermined target image. To support manual image mosaic creation, we have developed a software tool in which an image mosaic of a predetermined target image can be created through use of simple drag-and-drop functions. We refer to this as the *Mosaic Test Tool*. There are several features of the Mosaic Test Tool that have been specifically designed to simplify the process of manually generating an image mosaic. The Mosaic Test Tool is displayed simultaneously with the colour-based image retrieval system under evaluation.

As can be seen in Fig. 4, the Mosaic Test Tool is displayed on the left of the screen, using 30% of the available screen space. The colour-based image retrieval system being tested is then displayed in the remaining 70% of the screen. This removes the need for users to constantly switch between application windows, and permits users to easily drag images from the colour-based image retrieval system to their image mosaic in the Mosaic Test Tool.

The target image (the image the user is trying to replicate in the form of an image mosaic) and image mosaic under construction are displayed simultaneously in the Mosaic Test Tool interface to allow users to manually inspect and identify the colours (and colour layout) contained within each target image cell. As can be seen in Fig. 4, the target image is displayed in the top half of the Mosaic Test Tool. This is so the target image can act, much like the picture on a jigsaw-puzzle box, as a guide to completing the task. In the lower half of the Mosaic Test Tool is the image mosaic under construction, comprising the target image (at reduced opacity) overlaid with a grid. The reduced opacity target image underlay is designed to act as a guide for identifying the layout of the colours required in a database image to suitably fill an image mosaic cell. To fully inspect the actual colours that are required in a suitable database image, users can place the mouse cursor over a cell in the image mosaic. This highlights the corresponding target image section in the full colour target image, displayed in the top half of the Mosaic Test Tool.

Once users of the Mosaic Test Tool have located an image in the database (using the colour-based image retrieval system being evaluated) that they believe to be suitable to fill an image mosaic cell, they can drag the identified image from the

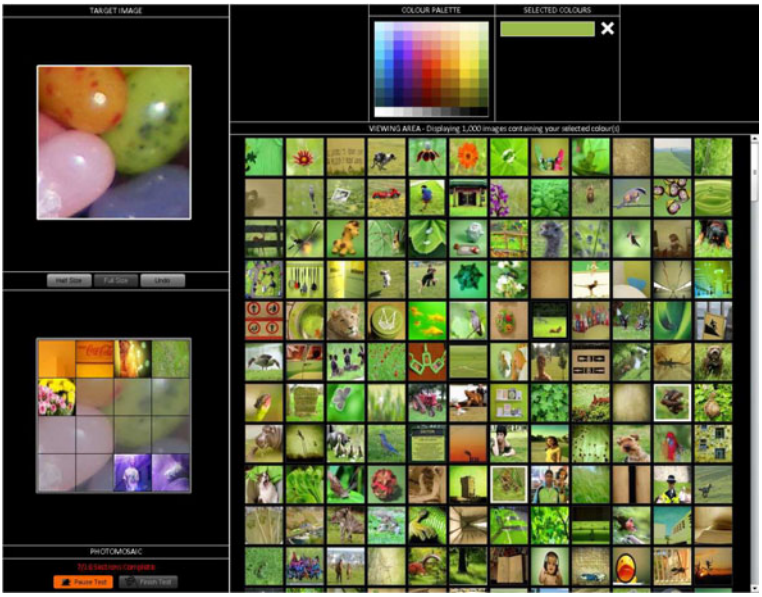

Fig. 4 A screenshot of the Mosaic Test Tool (*left*) and a colour-based image retrieval system (*right*) under evaluation during a Mosaic Test session. The 5 cells across the *top*, and the 2 cells in the *bottom right* hand corner, of the image mosaic have been filled with images from the database

colour-based image retrieval system directly to the desired image mosaic cell in the Mosaic Test Tool. It is important to note that the facility to export images through drag-and-drop operations is the only requirement of a colour-based image retrieval system for it to be compatible with the Mosaic Test Tool. If, upon reflection, users decide that the currently selected image is not suitable, however, they can simply drag the image out of the image mosaic cell, or revert to an earlier image via the 'undo' button

Located at the bottom of the Mosaic Test Tool interface are the 'Pause Test' and 'Finish Test' buttons. Since the time taken to complete image mosaics is an important measure of the effectiveness of the system under evaluation, the Mosaic Test can be easily paused should participants of a Mosaic Test require a break for any reason—thus preventing inappropriately extending task completion times. To prevent users submitting an incomplete image mosaic, the 'Finish Test' button is only enabled once all image mosaic cells are filled. When participants have submitted their image mosaic, the Mosaic Test Tool automatically records the total time taken as well as a bitmap of the user-generated image mosaic.

3.4 Target images

Photographs of jelly beans were used in the trial of the Mosaic Test. Not only do the images of jelly beans create a bright, interesting target image for participants to create in image mosaic form, but in addition it is possible for users to generate an image mosaic appearing visually similar to the target image (as can be seen in Fig. 5). During a pilot study of the Mosaic Test Tool, subsections of famous works

Fig. 5 An example of a jelly bean photograph target image (*left*) recreated by a Mosaic Test participant as a 4 × 4 image mosaic (*right*)

of art and major world landmarks were trialled as potential target images. It was found that participants had great difficulty in recreating such target images as image mosaics. As well as containing areas of intricate detail, the target images also had areas in which there are only subtle changes in colour (e.g. skin tones in paintings of faces). These slight differences in colour were mostly disregarded by the participants of the trial, resulting in the production of inaccurate and unconvincing image mosaics. Photographs of jelly beans, however, provide large areas of distinct colours, thus overcoming the difficulties experienced by participants in the study.

3.5 Training

Since the task is heavily reliant on colour matching, it is imperative that participants are tested for, or at least asked to self-report, any colour vision deficiencies (e.g., an inability to distinguish between 2 or more colours) before commencing a Mosaic Test. Participants are given written instructions explaining the concept of an image mosaic and the functionality of the Mosaic Test Tool. A practice session is undertaken by each participant, in which they are asked to complete a sample image mosaic using a small selection of suitable images. This is to ensure that all participants are trained to the same level in using the Mosaic Test Tool before commencing the test. Once participants are familiar with the functionality of the Mosaic Test Tool, and the evaluator has observed each participant complete a set of training tasks (such as dragging and removing images from the colour-based image retrieval system to the Mosaic Test Tool image mosaic), the participant can proceed to the measured Mosaic Test.

3.6 Comparing colour-based image retrieval systems

To demonstrate how 2 colour-based image retrieval systems can be directly compared using the Mosaic Test (and associated Mosaic Test Tool), we present a simple

case-study. A test user was asked to generate a 16 cell (4×4) image mosaic of a target image containing jelly beans (shown on the left of Fig. 5), using the Bing Images [9] and Google Images [3] internet search engines respectively. First, the user was trained on the functionality of the Mosaic Test Tool using written instructions. The test user was then shown how to use the colour-based image retrieval facility for filtering image search results in both systems. With Bing Images, users can select multiple colours from a graphical colour palette, located on the left hand side of the interface. For Google Images, colour-based image retrieval is performed by selecting a single colour from a graphical colour palette, also located on the left of the interface. Once briefed on using a system, the test user proceeded to the measured test. Here, she was required to complete her image mosaic using image results for the search terms *"flowers site: flickr.com"* (i.e., all images indexed by the respective systems with the keyword "flowers" from the online photo sharing community web site Flickr [23]). Upon completing an image mosaic with a system, the test user completed the NASA-TLX (task load index) [4] assessment.

As can be seen in Table 1, the Mosaic Test time, overall workload and relevance measurements achieved by the 2 systems in the case-study show that our test user was able to create an image mosaic using Google Images faster, more accurately and with less overall workload than when using Bing Images. At first glance these results suggest that for our test user at least, Google Images [3] offers more effective colour-based image retrieval than Bing Images. Of course, this is a single user case-study and any meaningful comparison of colour-based image retrieval systems using the Mosaic Test should be performed using as many participants as possible. It is also important to note that in this case-study, the test user created an image mosaic with Bing Images prior to making her image mosaic with the seemingly more effective Google Images. To alleviate any form of learning effect when undertaking a comparison of colour-based image retrieval systems using the Mosaic Test, the order in which participants are presented with the colour-based image retrieval systems under evaluation should be rotated evenly.

In this case-study, Google Images achieved lower measures than Bing Images for all 3 of the elements that comprise our definition of colour-based image retrieval effectiveness, making it straight forward to claim that Google Images provides more effective colour-based image retrieval than its Bing counterpart. It is indeed possible, however, when comparing colour-based image retrieval systems using the Mosaic Test that there may not be a system which is superior to others tested according to all 3 elements of our colour-based image retrieval effectiveness definition. For example, let us imagine that the test user in our case-study created an image mosaic

Table 1 The Mosaic Test measurements achieved by the Bing images [9] and Google images [3] internet search engines by the user in the case-study

System	Time (s)	Overall workload	Relevance
Bing images	1,054	9.67	3,147
Google images	902	6.83	2,953

Here, "*relevance*" represents the measured distance between the user image mosaic and initial target image (described further in Section 4.1)

faster and with less workload with Google Images, but less accurately than the image mosaic generated using Bing Images. Such a result would suggest that the colour-based image retrieval facility of Google Images is better suited to users who require images quickly, with a lesser regard for relevance. Applied to media production, if for example the Google Images system was used to index keyframes extracted from videos found online, this result scenario would suggest that Google Images would be more suitable for use in more time-constrained projects (e.g. finding suitably coloured graphics for use in a news article due to be aired during the next television news bulletin). This hypothetical result would also suggest that the colour-based image retrieval facility of Bing Images would be better suited for users who require the most relevant images to their requirements, but have more time available to retrieve them. Applying this to media production once more, Bing Images (in our hypothetical result) would be more suitable than Google Images for a graphic artist seeking realistic images for CGI to be used in a blockbuster film.

4 User study

To evaluate the reliability and suitability of the Mosaic Test for evaluating colour-based image retrieval systems, we recruited 24 users to participate in a user study. Of the 24 participants, 12 had previous experience working in graphic design, another creative industry which often requires people to retrieve images from a database on the basis of colour. In this *designer* group, 7 participants were male and 5 female. For our *non-designer* group, 10 participants were male and 2 female. All participants in the study reported no known colour-blindness or colour vision deficiency. Participants were asked to complete 3 image mosaics using 3 different colour-based image retrieval systems. Each of the colour-based image retrieval systems indexed the same image database, namely the 25,000 images contained within the MIRFLICKR-25000 collection [6]. The Mosaic Test Tool and colour-based image retrieval systems used were, for each participant, run on a Sony VAIO laptop, running Windows Vista, with a 17-inch (1600×900 resolution) display. This was to ensure that the colours displayed to users remained constant (as rendered colours can vary between graphic card and monitor manufacturers [8]).

Participants were first trained on the functionality of the Mosaic Test Tool using written instructions (as described earlier in Section 3.5). For each of the 3 colour-based image retrieval systems used, participants were first trained, and given an opportunity to practise with, the functionality of the system. Once users indicated to the evaluator that they were satisfied with the controls of the colour-based image retrieval system, they began the measured test session. Using each system, participants were asked to complete an image mosaic (comprising 16 cells, as per Fig. 5) of a different target image (all of jelly beans). This was to prevent users learning a set of suitable database images to use in a single image mosaic. The 3 target images were selected so that the number of jelly beans (and thus colours) in each were evenly balanced, with only the colour and layout of the jelly beans varying between the target images. To also ensure that results were not affected by one image mosaic being more difficult to complete than another, the order in which the target images

were presented remained constant whilst the colour-based image retrieval system order was counterbalanced across participants to guard against learning effects. After completing an image mosaic with a colour-based image retrieval system, participants were asked to complete a NASA-TLX (task load index) [4] assessment for the system they had just used.

The aim of the study was to investigate 3 primary factors relating specifically to the Mosaic Test as a method for reliably evaluating the effectiveness of colour-based image retrieval systems. Firstly, we hypothesised that users in the study would perform reflection-in-action and so we wanted to observe whether this was indeed true for participants when judging the suitability of images retrieved from the database. Secondly, in existing content-based image retrieval research, systems have been evaluated (using one of the evaluation methods described in Section 2.3) by recruiting either 'expert' [14] or 'non-expert' [13] users. We wanted to investigate what effect expert users or non-experts had on the time, workload and relevance data obtained from a Mosaic Test. We hypothesised that the expert users (graphic designers) would create more visually accurate image mosaics in less time and with less workload than the non-expert users, on account of the fact that graphic designers perform image retrieval for creative projects on a regular (if not daily) basis. Finally, we wanted to examine how well several image colour descriptors (and their associated distance measures) used in content-based image retrieval, correlate with *human* perceptions of image mosaic distance (i.e., the 'closeness' of an image mosaic compared with the target image). For this, after completing their 3 image mosaics, participants were asked to rank each of the submissions in ascending order of 'closeness' to its corresponding target image.

4.1 Measuring image relevance

Since an image mosaic is an art form intended to be viewed and enjoyed by humans, it seems logical that the adopted measure of image mosaic distance—i.e., how close an image mosaic is to its intended target image—should correlate with the inter-image distance perceptions of humans. An existing measure for automatically computing the distance between an image mosaic and its corresponding target image is the *Average Pixel-to-Pixel* (APP) distance [10]. The APP distance is expressed formally in (5), where n is the number of pixels in the mosaic image M and target image T, and r, g and b are the red, green and blue colour values of a pixel.

$$APP = \frac{\sum_{i=1}^{n} \sqrt{(r^i_M - r^i_T)^2 + (g^i_M - g^i_T)^2 + (b^i_M - b^i_T)^2}}{n} \tag{5}$$

Whilst Nakade and Karule [10] adopt the APP distance for calculating the accuracy of image mosaic algorithms, no research exists verifying that their adopted method is indeed a reliable approach to measuring the visual quality of an image mosaic. We therefore wanted to compare the existing APP image mosaic distance measure with a variety of image colour descriptors (and associated distance measures) commonly used in the domain of colour-based image retrieval in order to discover which correlates best with the human perception of image mosaic distance. To do this, we calculated the image mosaic distance rankings according to the existing

measure and several colour descriptors (and their associated distance measures), and then calculated the Spearman's rank correlation coefficient (r_s—used to measure the strength of a link between 2 sets of data) between each of the tested colour descriptor/distance measure combinations and the rankings assigned by the users in our study. We report only on colour descriptor and distance measure combinations which achieved a Spearman's rank correlation coefficient greater than 0.1.

For the image colour descriptors (and associated distance measures), we tested the global colour histogram (GCH) as an image descriptor [19]. We used a 64-bin histogram, in which each of the red, green and blue colour channels (in an RGB colour space) were quantised to 4 bins ($4 \times 4 \times 4 = 64$). We adopted the L_2 distance metric for comparing the global colour histograms of the image mosaics and corresponding target images. We also tested local colour histograms (LCH) as an image descriptor. For this, 64-bin colour histograms were also calculated for each image mosaic cell and the corresponding region in the target image (for the target image descriptor). The average Euclidean distance between all of the corresponding colour histograms (in the image mosaic and target image LCH descriptors) was used to compare LCH descriptors. Finally, we tested (along with their associated distance measures) the auto colour-correlogram descriptor [5], and the MPEG-7 colour-structure and colour-layout descriptors [17].

5 Results and discussion

5.1 Image relevance measures

Table 2 shows the Spearman's rank correlation coefficients (r_s) calculated between the human-assigned rankings and each of the rankings generated by the tested colour descriptor and distance measure combinations. The Spearman's rank correlation coefficient measures the strength of a link between 2 sets of data. We were interested in discovering which of the automatically generated rankings is most strongly linked to the human assigned rankings (i.e., human perception). We compared the correlation coefficient for each measure tested with the critical value, which at a 5% significance level with 22 degrees of freedom (24 participants - 2) equates to **0.423**. Any r_s value greater than this critical value can be considered a significant correlation at a 5% level.

As shown in Table 2, the MPEG-7 colour structure descriptor (MPEG-7 CST) was the only colour descriptor and associated distance measure we found to correlate

Table 2 The Spearman's rank correlation coefficients (r_s) between the image mosaic distance rankings made by humans and the rankings generated by the tested colour descriptors

Accuracy Measure	r_s	Significant (5%)
MPEG-7 CST	0.576	Yes
APP	0.266	No
GCH	0.255	No
MPEG-7 CL	0.188	No
LCH	0.166	No
ACC	0.144	No

with human perceptions of image mosaic distance at the 5% significance level. In other words, the MPEG-7 colour structure descriptor was the only measure which identified relevance that resonated with the human assigned relevance. Therefore, by measuring the L_1 distance between the MPEG-7 CST of the target and user-generated image mosaics, the Mosaic Test can automatically calculate the relevance of retrieved images according to the low-level feature of colour in a manner that correlates with human perception. This validated relevance measure addresses the first drawback of current evaluation methods.

5.2 Reflection-in-action observations

As part of our user study, we observed the actions performed by participants when creating an image mosaic. It was clear that the majority of users, in both the designer and non-designer groups, did perform reflection-in-action [16] when assessing the relevance of images retrieved from the database. The manner in which reflection-in-action was physically realised, however, varied between users. Some users relied on the 'undo' button: to assess the greater potential suitability of an image from the database relative to a pre-existing image in an image mosaic cell, users would overwrite the pre-existing image with the newly retrieve image. If the newly retrieved image was less suitable than the pre-existing image, users would click 'undo' to revert back to the pre-existing image. This observed behaviour corresponds with similar 'undo'-based reflection-in-action as witnessed amongst creative individuals by Terry and Mynatt [20]. Another popular reflection-in-action strategy across users in both groups was to drag and 'hover' a retrieved image from the tested colour-based image retrieval system over the intended image mosaic cell to inspect its suitability. Irrespective of their chosen enactment of reflection-in-action, from the very fact that participants were observed performing reflection-in-action, it is clear that the Mosaic Test overcomes the second of the 2 major drawbacks occurring in existing evaluation methods.

5.3 Comparing designer and non-designer performance

To compare the performance of graphic designers and non-designers participating in our study, we analysed each of the colour-based image retrieval effectiveness elements, namely task completion time (measured in seconds), user workload (measured using the mean rating of the 6 NASA-TLX scales) and relevance (assessed using MPEG-7 colour structure descriptors) for both groups. All significance tests in this section were carried out using a one-way ANOVA with a 5% significance level.

5.3.1 Workload

The first element of our colour-based image retrieval effectiveness definition is effort. We hypothesised that the perceived effort expended by graphic designers would be lower than that of non-designers. Figure 6 shows that, as expected, the average perceived effort expended by graphic designers is lower than that of non-designers. Interestingly, however, this difference was found not to be significant ($F_{1,70} = 0.49$, $p > 0.05$).

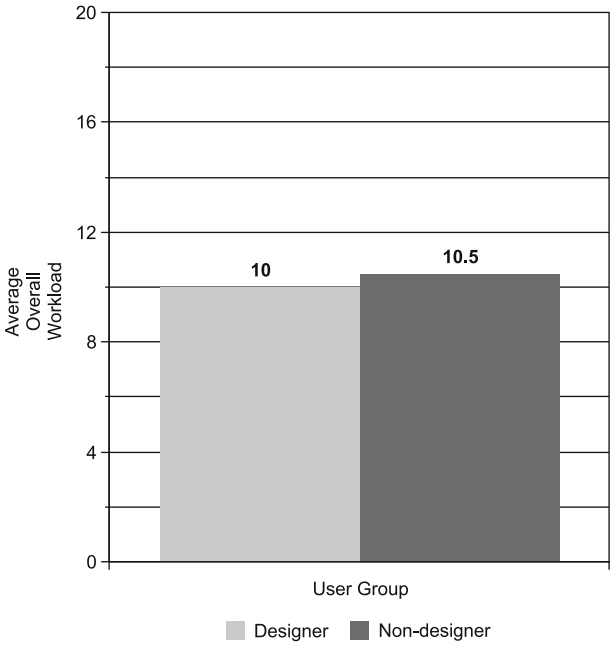

Fig. 6 The mean *"overall workload"* experienced by users in the designer and non-designer groups

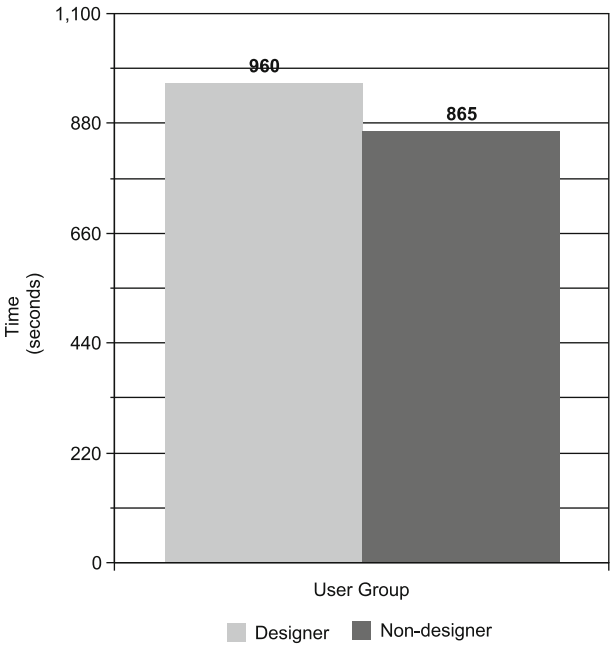

Fig. 7 The mean time (in seconds) required by users in the designer and non-designer groups to complete an image mosaic

5.3.2 Time

The second of the 3 elements in our colour-based image retrieval effectiveness definition is time. We hypothesised that graphic designers would require less time to complete their image mosaics than non-designers. We were surprised to observe, however, that (on average) non-designers completed their image mosaics faster than the graphic designers. As can be seen in Fig. 7, the graphic designers required an average of 95 seconds longer than non-designers to complete their image mosaics. Upon observing the graphic designers creating their image mosaics, it became apparent that some continually replaced (i.e., attempted to upgrade) the images that they had already assigned to image mosaic cells—that is, they engaged in more extended reflection-in-action. The decision to replace such images would often occur once all cells of the image mosaic had been assigned an image. It is important to note that the differences in time required for completing an image mosaic were not found to be significant ($F_{1,70} = 0.52$, $p > 0.05$).

5.3.3 Relevance

The final element of our colour-based image retrieval effectiveness definition is relevance. As discussed in Sections 4.1 and 5, the relevance of images retrieved during a Mosaic Test is measured automatically by calculating the L_1 distance between the MPEG-7 colour structure descriptors of user-generated image mosaics and the corresponding target image. We hypothesised that graphic designers would create image mosaics that were visually closer to the initial target image. As illustrated by Fig. 8, however, the average image mosaic distance of non-designers was lower

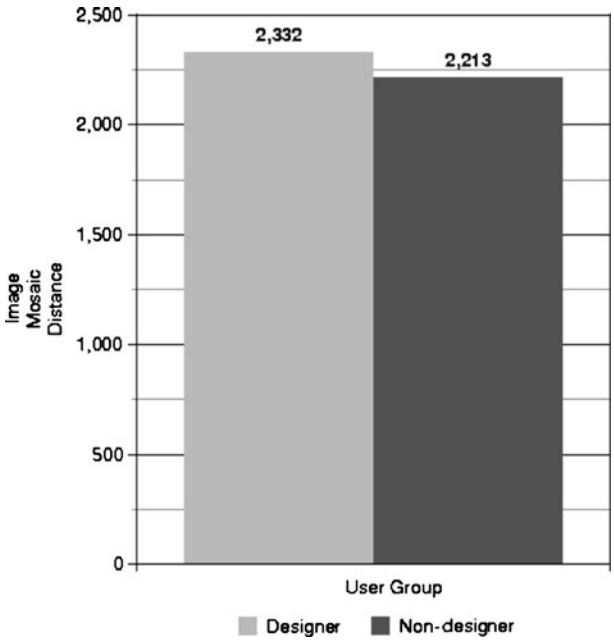

Fig. 8 The mean L_1 distance between the MPEG-7 CSTs of image mosaics and target images created by users in the designer and non-designer groups

than that of the designers. We did not find this difference to be significant ($F_{1,70} = 1.46$, $p > 0.05$).

5.3.4 Designer vs. non-designer discussion

The results of our user study show that there is no significant difference between the time, workload and relevance measures achieved by 'expert' and 'non-expert' participants. We can therefore reject our hypothesis that 'expert' users (graphic designers) create more visually accurate image mosaics in less time and incur less workload than 'non-expert' users. As a result, to reliably evaluate the effectiveness of colour-based image retrieval systems using the Mosaic Test, it is not necessary to recruit 'expert' users for testing—often an expensive, difficult to plan and most of all time-consuming task. Instead, a sample population of computer-literate participants can be tested, thus overcoming the difficulties of recruiting 'expert' users. This is another clear advantage of the Mosaic Test.

6 Conclusion

A variety of content-based image retrieval systems exist which enable users to perform image retrieval based on colour content—i.e., colour-based image retrieval. For the production of media for use in television and film, colour-based image retrieval is useful for retrieving specifically coloured animations, graphics or videos from large databases (by comparing user queries to the colour content of extracted key frames). It is also useful to graphic artists creating realistic computer-generated imagery (CGI). Unfortunately, current methods for evaluating colour-based image retrieval systems have 2 major drawbacks. Firstly, the *relevance* of images retrieved during the task cannot be measured reliably. Secondly, existing methods do not account for the creative design activity known as *reflection-in-action*. Consequently, the development and application of novel and potentially more effective colour-based image retrieval approaches, better supporting the large number of users creating media for use in television and film productions, is not possible as their efficacy cannot be reliably measured and compared to existing technologies. In this research, we have introduced the Mosaic Test which has been developed to address this problem by providing a reliable mechanism by which to meaningfully evaluate colour-based image retrieval systems.

The findings of a user study, in which we evaluated the Mosaic Test using 24 participants, have confirmed that the Mosaic Test overcomes the 2 major drawbacks associated with previous evaluation methods: in addition to providing valuable effectiveness data relating to efficiency and user effort, the Mosaic Test enables participants to reflect on the relevance of retrieved images within the context of their image mosaic (i.e., to perform reflection-in-action [16]), and automatically measures the relevance of retrieved images, in a manner which correlates with human perception of relevance by computing MPEG-7 colour structure descriptors (from the user-generated image mosaics and target images) and calculating the L_1 distance between them. The results of our user study also show that the participants need not be 'expert' colour-based image retrieval system users in order to reliably evaluate

the effectiveness of colour-based image retrieval systems using a Mosaic Test. This is important as it removes difficulties, such as time and finance, often associated with recruiting expert users for software testing.

As a result of these findings, we propose that the Mosaic Test be adopted in all future research and practice evaluating and comparing the effectiveness of colour-based image retrieval systems. To this end, we will be publicly releasing the Mosaic Test Tool and procedural documentation for other researchers in the domain of colour-based image retrieval.

Acknowledgements We would like to sincerely thank everyone that participated in the user study described in this work, all of which gave up their spare time for no financial award. This work has been funded by the Aston University School of Engineering and Applied Science's EPSRC doctoral training grant.

References

1. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
2. Faloutsos C, Equitz W, Flickner M, Niblack W, Petkovic D, Barber R (1994) Efficient and effective querying by image content. J Intell Inf Syst 3:231–262
3. Google (2010) Google images. http://images.google.com/. Accessed 2 Nov 2010
4. Hart SG (2006) NASA-Task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society 50th annual meeting, pp 904–908
5. Huang J, Kumar SR, Mitra M, Zhu W, Zabih R (1997) Image indexing using color correlograms. In: Computer vision and pattern recognition, pp 762–768
6. Huiskes MJ, Lew MS (2008) The MIR flickr retrieval evaluation. In: ACM international conference on multimedia information retrieval, pp 39–43
7. iStockPhoto.com (2010) http://www.istockphoto.com/. Accessed 1 Dec 2010
8. MacDonald L (1999). Using color effectively in computer graphics. IEEE Comput Graph Appl 19(4):20–35
9. Microsoft (2011) Bing images. http://www.bing.com/images. Accessed 27 Jan 2011
10. Nakade S, Karule P (2007) Mosaicture: image mosaic generating system using CBIR technique. In: International conference on computational intelligence and multimedia applications, pp 339–343
11. Ortega M, Rui Y, Chakrabarti K, Mehrotra S, Huang TS (1997) Supporting similarity queries in MARS. In: ACM international multimedia conference, pp 403–413
12. Plant W, Schaefer G (2009) Evaluation and benchmarking of image database navigation tools. In: International conference on image processing, computer vision, and pattern recognition, pp 248–254
13. Rodden K, Wood K (2003) How do people manage their digital photographs? In: SIGCHI conference on human factors in computing systems, pp 409–416
14. Rodden K, Basalaj W, Sinclair D, Wood K (2001) Does organisation by similarity assist image browsing? In: SIGCHI conference on human factors in computing systems, pp 190–197
15. Schaefer G, Stich M (2004) UCID—an uncompressed colour image database. In: Storage and retrieval methods and applications for multimedia, pp 472–480
16. Schön DA (1983) The reflective practitioner: how professionals think in action. Basic Books
17. Sikora T (2001) The MPEG-7 visual standard for content description—an overview. IEEE Trans. Circuits Syst Video Technol 11(6):696–702
18. Silvers R (1996) Photomosaics: putting pictures in their place. Master's thesis, Massachusetts Institute of Technology
19. Swain M, Ballard D (1991) Color indexing. Int J Comput Vis 7(1):11–32
20. Terry M, Mynatt ED (2002) Recognizing creative needs in user interface design. In: Creativity and cognition, pp 38–44

21. Truong BT, Venkates S (2007) Video abstraction: a systematic review and classification. ACM Trans Multimedia Comput Commun Appl 3(1):1–37
22. Vijfwinkel M (2009) CG textures. http://www.cgtextures.com/. Accessed Oct 2009
23. Yahoo (2009) Flickr. http://www.flickr.com/. Accessed Oct 2009
24. Zhang Y, Nascimento M, Zaiane O (2003) Building image mosaics: an application of content-based image retrieval. In: International conference on multimedia and expo, pp 317–320

William Plant recieved his BSc in Computer Science in 2008 from Aston University, UK. He is currently studying towards a Ph.D with the School of Engineering and Applied Science at Aston University. His current research is focussed on effective and efficient techniques for retrieving images from large databases according to colour content. His other research interests include image database visualisation and browsing.

Joanna Lumsden received her BSc in Software Engineering in 1996 and a Ph.D. in Human Computer Interaction (HCI) in 2001, both from Glasgow University, Scotland. She is currently a lecturer and Manager of the Aston Interactive Media (AIM) Lab at Aston University. Her research activities cover many aspects of human computer interaction. Most recently she has been focussing on trust in e- Commerce and on mobile human computer interaction design and associated evaluation techniques. Joanna has a keen interest in using mobile technologies in novel capacities (e.g., as

assistive devices) and in designing mobile applications to best support field operatives. Joanna is the Editor-in-Chief of the International Journal of Mobile HCI (IJMHCI), the only journal dedicated to the design and evaluation of the user interface aspects of mobile technologies. She also served as the editor of the Handbook of Research on User Interface Design and Evaluation for Mobile Technology which brought together more than 60 quality contributions from over 100 world renowned experts in the field of mobile HCI.

Ian T. Nabney studied mathematics at Oxford and Cambridge Universities before joining Logica's research labs where he spent five years working in neural networks and compiler verification. He joined Aston University in 1995, where he is now Professor of Computer Science. He is the lead developer of the Netlab pattern analysis toolbox, which has more than 45,000 users, both academic and industrial, worldwide. He has published more than 75 peer-reviewed publications in both the theory and a diverse range of applications of machine learning, with a particular focus on data visualisation methods and time series analysis.