The Mosaic Test: Benchmarking Colour-based Image Retrieval Systems Using Image Mosaics

William Plant
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

Joanna Lumsden
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

Ian T. Nabney
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

ABSTRACT

Evaluation and benchmarking in content-based image retrieval has always been a somewhat neglected research area, making it difficult to judge the efficacy of many presented approaches. In this paper we investigate the issue of benchmarking for *colour-based image retrieval* systems, which enable users to retrieve images from a database based on low-level colour content alone. We argue that current image retrieval evaluation methods are not suited to benchmarking colour-based image retrieval systems, due in main to not allowing users to reflect upon the suitability of retrieved images within the context of a creative project and their reliance on highly subjective ground-truths. As a solution to these issues, the research presented here introduces the *Mosaic Test* for evaluating colour-based image retrieval systems, in which test-users are asked to create an image mosaic of a predetermined target image, using the colour-based image retrieval system that is being evaluated. We report on our findings from a user study which suggests that the Mosaic Test overcomes the major drawbacks associated with existing image retrieval evaluation methods, by enabling users to reflect upon image selections and automatically measuring image relevance in a way that correlates with the perception of many human assessors. We therefore propose that the Mosaic Test be adopted as a standardised benchmark for evaluating and comparing colour-based image retrieval systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*; H.2.8 [**Database Management**]: Database Applications—*Image Databases*

Keywords

Image databases, content-based image retrieval, image mosaic, performance evaluation, benchmarking.

1. INTRODUCTION

Colour-based image retrieval systems such as Chromatik [1], MultiColr [5] and Picitup [10] enable users to retrieve images from a database based on colour content alone. Such a facility is particularly useful to users across a number of different creative industries, such as graphic, interior and fashion design [6, 7]. Surprisingly, however, little research appears to have been conducted into evaluating colour-based image retrieval systems. Currently, there is no standardised measure and image database to evaluate the performance of an image retrieval system [8]. The most commonly applied evaluation methods are those of *precision and recall* [8] and the *target search* and *category search* tasks [11]. The precision and recall measure is used to evaluate the accuracy of image results returned by a system in response to a query, whilst the target search and category search tasks are both user-based evaluation strategies in which test-users are asked to retrieve images from a database that are relevant to a given target, using the image retrieval system that is being evaluated.

In this research, we argue that the image retrieval system evaluation strategies listed above are not suitable for evaluating and benchmarking colour-based image systems for two fundamental reasons. Firstly, none of the above evaluation methods allow test-users to perform an important process often conducted by creative users, known as *reflection-in-action* [12]. In reflection-in-action, a creative project is modified by a user and then reviewed by the user after the modification. After assessing their modification, the creative individual will then decide whether to maintain or discard the modification to the project. As an example, a graphic designer will add an image to a web page before making an assessment as to its aesthetic suitability. Secondly, the category search and precision and recall measures require an image database and associated ground-truth (a manually generated list pre-defining which images in the database are similar to others) for defining image relevance during a system evaluation. Such human-based definitions of similarity, however, can often be highly subjective resulting in retrieved images being incorrectly assessed as irrelevant.

As a result of these drawbacks, no method currently exists for reliably evaluating colour-based image retrieval systems. The following section introduces the Mosaic Test which has been developed to address the current problem, providing a reliable means for benchmarking colour-based image retrieval systems.

2. THE MOSAIC TEST

For the Mosaic Test, participants are asked to manually create an image mosaic (comprising 16 cells) of a predetermined target image. An image mosaic (first devised by Silvers [14]) is a form of art that is typically generated automatically through use of content-based image analysis. A target image is divided into cells, each of which is then replaced by a small image with similar colour content to the corresponding cell in the target image. Viewed from a distance, the smaller images collectively appear to form the target image, whilst viewing an image mosaic close up reveals the detail contained within each of the smaller images. An example of an automatically generated image mosaic is shown in Figure 1.

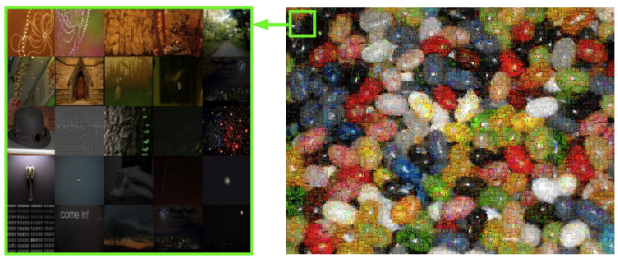

Figure 1: An example of an image mosaic. The region highlighted green in the image mosaic (right) has been created using the images shown (left).

For target images in the Mosaic Test, photographs of jelly beans are used. The images of jelly beans produce a bright, interesting target image for participants to create in mosaic form and the generation of an image mosaic that appears visually similar to the target image is also very achievable. More importantly, retrieving images from a database comprising large areas of a small number of distinct colours is a practise commonly performed by users in creative industries.

To complete their image mosaics, participants must identify the colours required to fill an image mosaic cell (by inspecting the corresponding region in the target image), and retrieve a suitably coloured image from the 25,000 contained within the MIRFLICKR-25000 image collection [4] using the colour-based evaluation system under evaluation. When selecting images for use in their image mosaic, users can add, move or remove images accordingly to assess the suitability of images within the context of their image mosaic. It is in this way that the Mosaic Test overcomes the first major drawback of existing evaluation methods, by enabling participants to perform the creative practise of reflection-in-action [12]. Upon completion of an image mosaic, the time required by the user to finish the image mosaic is recorded, along with the visual accuracy of their creation in comparison with the initial target image. Through analysing the accuracy of user-generated image mosaics (in a manner which correlates with the perception of a number of different human assessors), the Mosaic Test is able to overcome the second drawback associated with existing evaluation techniques. This is because it does not rely on a highly subjective image database ground-truth. The image mosaic accuracy measure adopted for use with the Mosaic Test is discussed further in Section 3.1. Additionally, participants are asked to indicate their subjective experience of workload (using the NASA TLX scales [2]) post test.

The time (number of seconds), subjective workload (user NASA-TLX ratings) and relevance (image mosaic accuracy) measures achieved by colour-based image retrieval systems evaluated using the Mosaic Test can be directly compared and used for benchmarking. When comparing the Mosaic Test measures achieved by different systems, the more effective colour-based image retrieval system will be the one that enables users to create the most accurate image mosaics, fastest and with the least workload.

2.1 Mosaic Test Tool

To support users in their manual creation of image mosaics using the Mosaic Test, we have developed a novel software tool in which an image mosaic of a predetermined target image can be created using simple drag and drop functions. We refer to this as the *Mosaic Test Tool*. The Mosaic Test Tool has been designed so that it can be displayed simultaneously with the colour-based image retrieval system under evaluation (as can be seen in Figure 2). 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 being tested to their image mosaic 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 and thus the Mosaic Test.

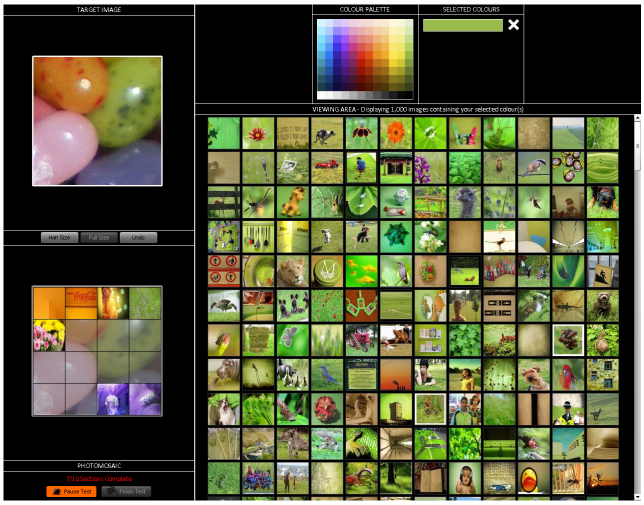

Figure 2: The Mosaic Test Tool (left) and an image retrieval system under evaluation (right) during a Mosaic Test session.

The target image and image mosaic are displayed simultaneously on the Mosaic Test Tool interface to allow users to manually inspect and identify the colours (and colour layout) required for each image mosaic cell. As can be seen in Figure 2, the target image (the image the user is trying to replicate in the form of an image mosaic) is displayed in the top half of the Mosaic Test Tool. Coupled with the ease in which images can be added to, or removed from, image mosaic cells, users of the Mosaic Test Tool can simply as-

sess the suitability of a retrieved image by dragging it to the appropriate image mosaic cell and viewing it alongside the other image mosaic cells.

3. USER STUDY

To evaluate the Mosaic Test, we recruited 24 users to participate in a user study. Participants were given written instructions explaining the concept of an image mosaic and the functionality of the Mosaic Test Tool. A practise session was undertaken by each participant, in which they were asked to complete a practise image mosaic using a small selection of suitable images. Participants were then asked to complete 3 image mosaics using 3 different colour-based image retrieval systems. To ensure that users did not simply learn a set of database images suitable for use in a solitary image mosaic, 3 different target images were used. These target images were carefully 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 effected by a target image being more difficult to create in image mosaic form than another, the order in which the target images were presented to participants remained constant whilst the order in which the colour-based image retrieval systems were used was counter balanced. After completing the 3 image mosaics, participants were asked to rank each of their creations in ascending order of 'closeness' to its corresponding target image.

We wanted to investigate whether the Mosaic Test does overcome the drawbacks of existing evaluation strategies so that it may be adopted as a reliable benchmark 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, we were eager to investigate which method should be adopted for measuring the accuracy of an image mosaic in the Mosaic Test.

3.1 Assessing Image Mosaic Accuracy

As 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 accuracy - i.e., how close an image mosaic looks to its intended target image - should correlate with the inter-image distance perceptions of a number of human assessors. 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 [9]. The APP distance is expressed formally in Equation (1), where i is 1 of a total n corresponding 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.

$$APP = \frac{\sum_{i=0}^{n} \sqrt{(r_M^i - r_T^i)^2 + (g_M^i - g_T^i)^2 + (b_M^i - b_T^i)^2}}{n} \quad (1)$$

We were eager to compare the existing APP image mosaic distance measure with a variety of image colour descriptors (and associated distance measures) commonly used for content-based image retrieval, to discover which best correlates with human perceptions 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 between each of the tested distance measures and the rankings assigned by the users in our study.

For the image colour descriptors (and associated distance measures), we firstly tested the global colour histogram (GCH) as an image descriptor. A colour histogram contains a normalised pixel count for each unique colour in the colour space. 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 x 4 x 4 = 64). We adopted the Euclidean distance metric to compare 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 calculated for each image mosaic cell (for the image mosaic descriptor), and its corresponding area 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 MPEG-7 colour structure (MPEG-7 CST) and colour layout (MPEG-7 CL) descriptors [13], as well as the auto colour correlogram descriptor (ACC) [3].

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), a sliding window (8×8 pixels in size) moves across the image in the HMMD colour space [13] (reduced to 256 colours). 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 two MPEG-7 CSTs or two ACCs can be calculated using the L_1 (or city-block) distance metric. Finally, the MPEG-7 colour layout descriptor (MPEG-7 CL) [13] divides an image into 64 regular blocks, and calculates the dominant colour of the pixels within each block [13]. The cumulative distance between the colours (in the YC_bC_r colour space) of corresponding blocks forms the measure of similarity between 2 MPEG-7 CL descriptors.

Accuracy Measure	$\mathbf{r_s}$	**Significant (5%)**
MPEG-7 CST	0.572	YES
APP	0.275	NO
GCH	0.242	NO
MPEG-7 CL	0.198	NO
LCH	0.176	NO
ACC	0.154	NO

Table 1: 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.

4. RESULTS

Table 1 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 descriptors. We compare the r_s correlation coefficient for each measure tested with the critical value of r, which at a 5% significance level with 22 d.f. $(24 - 2)$ equates to **0.423**. Any r_s value greater than this critical value can be considered a significant correlation at a 5% level.

5. DISCUSSION

We observed the actions taken by the participants of the user study when creating their image mosaics. It was clear that the majority of users performed reflection-in-action when assessing the relevance (or suitability) of images retrieved from the database for use in their image mosaics. As participants of a Mosaic Test were able to perform this reflection-in-action [12], it is clear that the Mosaic Test also overcomes the first of the two major drawbacks present in current image retrieval evaluation methods. As shown in Table 1, the MPEG-7 colour structure descriptor (MPEG-7 CST) was the only colour descriptor (and associated distance measure) we found to correlate with human perceptions of image mosaic distance at the 5% significance level. Therefore, by measuring the L_1 (or city-block) distance between the MPEG-7 CSTs of the target image and user-generated image mosaics, the Mosaic Test can automatically calculate the relevance of retrieved images in a manner that correlates with human perception, thus overcoming the second major drawback of existing image retrieval evaluation methods for benchmarking colour-based image retrieval systems (the reliance on a highly subjective image database ground-truth).

6. CONCLUSION

Current image retrieval system evaluation methods have two fundamental drawbacks that result in them being unsuitable for evaluating and benchmarking colour-based image retrieval systems. These evaluation strategies do not enable users to perform the practise of reflection-in-action [12], in which creative users assess project modifications within the context of the creative piece he/she is working on. The existing image retrieval system evaluation methods also rely heavily upon highly subjective image database ground-truths when assessing the relevance of images selected by test users or returned by a system. As a result of these drawbacks, no method currently exists for reliably evaluating and benchmarking colour-based image retrieval systems. In this paper, we have introduced the Mosaic Test which has been developed to address the current problem, by providing a reliable means by which to evaluate colour-based image retrieval systems.

The findings of a user study reveal that the Mosaic Test overcomes the two major drawbacks associated with existing evaluation method used in the research domain of image retrieval. As well as also providing valuable effectiveness data relating to efficiency and user workload, the Mosaic Test enables participants to reflect on the relevance of retrieved images within the context of their image mosaic (i.e., perform reflection-in-action [12]). The Mosaic Test is also able to automatically measure the relevance of retrieved images in a manner which correlates with the perceptions of multiple human assessors, by computing MPEG-7 colour structure descriptors from the user-generated image mosaics and their corresponding target images, and calculating the L_1 (or city-block) distance between them. As a result of our findings, we propose that the Mosaic Test be adopted in all future research evaluating the effectiveness of colour-based image retrieval systems. Future work will be to publicly release the Mosaic Test Tool and procedural documentation for other researchers in the domain of content-based image retrieval.

7. REFERENCES

[1] Exalead. Chromatik. Accessed December 1, 2010, at: `http://chromatik.labs.exalead.com/`.

[2] S. G. Hart. NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pages 904–908, 2006.

[3] J. Huang, S. R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In *Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[4] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *ACM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[5] idée Inc. idée MultiColr Search Lab. Accessed November 2, 2010 at `http://labs.ideeinc.com/multicolr`.

[6] Imagekind Inc. Shop Art by Color. Accessed November 2, 2010, at: `http://www.imagekind.com/shop/ColorPicker.aspx`.

[7] T. K. Lau and I. King. Montage : An Image Database for the Fashion, Textile, and Clothing Industry in Hong Kong. In *Third Asian Conference on Computer Vision*, pages 410–417, 1998.

[8] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[9] S. Nakade and P. Karule. Mosaicture: Image Mosaic Generating System Using CBIR Technique. In *International Conference on Computational Intelligence and Multimedia Applications*, pages 339–343, 2007.

[10] Picitup. Picitup. Accessed January 21, 2011, at: `http://www.picitup.com/`.

[11] W. Plant and G. Schaefer. Evaluation and Benchmarking of Image Database Navigation Tools. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, pages 248–254, 2009.

[12] D. A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, 1983.

[13] T. Sikora. The MPEG-7 Visual Standard for Content Description - An Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 2001.

[14] R. Silvers. Photomosaics: Putting Pictures in their Place. Master's thesis, Massachusetts Institute of Technology, 1996.