# A Multidimensional Query Distance Space for Image Retrieval by Colour

D. ANDROUTSOS
Electrical & Computer Engineering
Ryerson University
350 Victoria Street, Toronto, ON  M5B 2K3
CANADA

*Abstract:* - A Multidimensional Query Distance Space (MQDS) is introduced for image retrieval. This is a query-dependent space whose dimension is defined by the number of query colours. Specifically, the distances of the closest indexed colours to the query colours form a vector which lies in the MQDS. The location of this distance vector within the space and its relation to the origin and equidistant line, determines the overall ranking of a given image. The proposed scheme exhibits great flexibility. Querying can be performed in a number of ways, including query-by-colour and query-by-example and also allows incorporation of colour exclusion.

*Keywords:* - Image, retrieval, query, colour, multidimensional, vector, exclusion

## 1 Introduction

The most widely used approach to database retrieval has been modeling data and queries as vectors [1]. This is primarily due to the fact that some simple measures of similarity can be implemented easily and quickly, such as the $L_1$ metric. Typically, for a given query, a *query vector* is created and compared to all stored data vectors, (i.e., indices), to determine a measure of similarity. These similarity values are then sorted and all those which exhibit similarity below a redefined threshold are retrieved as the most similar. Alternatively, the first *k* most similar can be retrieved from the sorted set. Such techniques developed and evolved initially for text retrieval [2], but have made their way into image retrieval. This is because early colour image indices were nothing more than colour histograms [5]. Consequently, most systems today still subscribe to this type of scheme and are based on high-dimensional vector indexing and similarity. For example, most systems still implement colour histograms, which are essentially vectors whose dimensionality is determined by the number of histogram bins. The colours of an image are mapped onto a discrete colour space of *n* colours. Then, this *n* bin histogram forms an *n*-dimensional vector, which is compared to other computed histograms to determine a similarity value to each other. Unfortunately, the use of histograms for image retrieval via colour, is wrought with problems. For example the lack of spatial information and retrieval on a global colour scale are issues that affect retrieval results. Another not so obvious drawback, but as important, is the fact that by using histogram techniques there is no inherently effective procedure for *excluding* colours from a query.

In this paper, we do away with histogram indexing techniques and instead implement RGB vector techniques. This way we end up with a much smaller index which does not have the over-completeness or *granularity* of a colour histogram, yet retrieval performance is better and more robust. More specifically, for each database image, recursive HSV-space segmentation is performed to extract regions of prominent and perceptually relevant colour [3]. The number of extracted colours proves to be very low, yet very effective. The average RGB values of the extracted colours are used as representative image vectors which are stored and indexed, along with information regarding the amount of each colour present in each image, the number of regions which contain each colour, and the approximate colour category to which they belong.

To perform the actual image similarity we use a perceptually tuned vector-angular based similarity measure [4] on each extracted colour of each image being compared and introduce the Multidimensional Query Distance Space to combine all calculated values for an overall measure. By implementing this novel multidimensional vector space, we can perform quick image retrieval in a variety of ways, and in addition can perform effective queries by also specifying which colours to *exclude* from retrieved images.

## 2 Indexing

The first step in an image retrieval system is the generation of indices, i.e., feature extraction. Since we are dealing with the retrieval of images based on colour content, we aim to identify prominent colours in each database image. To this end, we apply a recursive HSV-space segmentation technique to extract colours. The HSV cone is partitioned into perceptual segments, which allows colours to be identified as white, black, chromatic and bright chromatic. This is then followed by a post-processing stage to remove small irrelevant objects and to identify all colour regions and their locations [3]. The RGB vectors of these segmented colours, or representative vectors, are then stored as indices in the database, to later be used for similarity calculation and ultimately image retrieval. We have found that in our entire 2000 test image database, the average number of colour regions extracted is 4.86 with a maximum of 16. Figure 1 shows an example of an original image and its segmented and post-processed output.

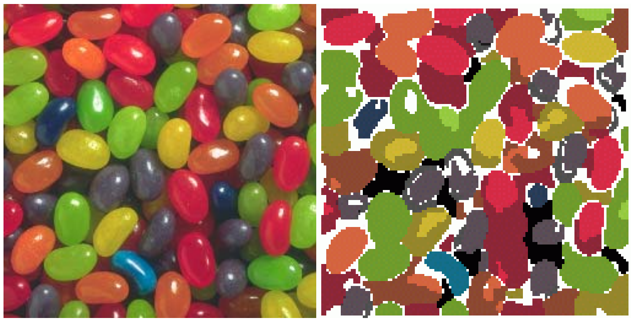


*Figure 1*: Sample image and its feature extraction results. *Left*: Original image, *right*: segmented and post-processed result.

The actual image retrieval is achieved by calculating a similarity value between the RGB vectors of user-defined query colours to the indexed colour vectors that have been extracted and stored. This similarity value is used to determine the retrieval ranking of an image in the database implying how similar it is to the query posed.

## 3 Multidimensional Query Distance Space (MQDS)

We propose a vector representation of the similarity values, which we refer to as the *multidimensional query distance vector* (MQDV) **D**, defined as:

$$\mathrm{D}(\mathrm{q}_1,\cdots,\mathrm{q}_n) = \mathrm{I} - (s_1, s_2, \cdots, s_n), \qquad (1)$$

where **q** are the *n* input 3-dimensional RGB query colours . These can be either user defined colours or colours extracted and determined via query-by-colour. **I** is a vector of size *n* with all entries set to 1 and $\{s_1, s_2, \ldots, s_n\}$ are the similarity values of each query colour to the extracted and indexed colour vectors $\{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_r\}$ of an image in the database, defined as:

$$s_j = \max(S(\mathrm{q_j}, \mathrm{i_1}), \cdots, S(\mathrm{q_j}, \mathrm{i_r})), \qquad (2)$$

where *S* represents any vector similarity calculation.

As an example, assume that a query consists of 2 query colours, $\mathbf{q}_1$and $\mathbf{q}_2$, and a given index contains pre-indexed colour vectors, $\mathbf{i}_1$ and $\mathbf{i}_2$. The maximum similarity $s_1$ between $\mathbf{q}_1$ and $\mathbf{i}_1$ and $\mathbf{i}_2$ is taken, along with the similarity $s_2$ between $\mathbf{q}_2$ and $\mathbf{i}_1$ and $\mathbf{i}_2$, to build $\mathbf{D}(\mathbf{q}_1, \mathbf{q}_2)$. The process is repeated for each pre-indexed database image, resulting with *M* distance vectors, where *M* is the total number of indices upon which similarity was calculated. The maximum value of *M* is, of course, the total number of images in the database. This set of vectors **D** span what we call the *multidimensional query distance space* (MQDS). For the case of 2 query colours, the space is two-dimensional, for 3 colours it is three-dimensional, etc. Each database image exists at a point in this space, by virtue of its distance vector **D,** and its location can be used to calculate a retrieval ranking, $\Re$, for the corresponding image based on the given query. More specifically, the magnitude and orientation of **D** determines *which* of the query colours the image index is most similar to, and also determines the *degree* of similarity to each of the query colours. For the special case of *single colour* query, where only one colour is specified in the database query, the MQDV is a one-dimensional vector and the MQDS reverts to a scalar comparison of this similarity value.

### 3.1 Equidistant Line

The key to calculating the image ranking lies in the origin of the MQDS and the *equidistant line*. The *equidistant line* is the set of all points in the MQDS where all component values of **D** are equal. Figure 2 depicts the MQDS. For example, in a 3-D MQDS, the e*quidistant line* passes through (0,0,0) and (1,1,1). In other words, all MQDVs along this line represent image indices whose representative vectors exhibit an equal amount of similarity to their corresponding query colours.

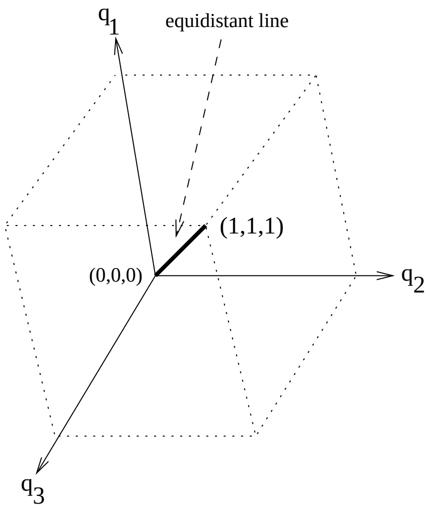


*Figure 2*: A visualization of the MQDS for 3 query colours, ($q_1$, $q_2$, and $q_3$), showing the equidistant line.

Clearly then, the database image that is the closest match to *all* the given query colours $\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_n$, is the one whose index produces an MQDV that is collinear with the *equidistant line* and at the same time is closest to the origin, (i.e., has the smallest magnitude). Figure 3 provides an intuitive representation to this concept. The location of each tiny image displayed in the two dimensional MQDS, formed by the two query colours *green* (RGB = 26,153,33) and *red* (RGB = 200,7,25), corresponds to each image's calculated **D**. How close an image is to each of the axes quantifies how similar that particular image's colour content is to the query colour which is represented by that axis; i.e., images that are closer to the *green* axis contain a colour that is closer in similarity to the *green* query colour, as compared to the similarity that one of its colours exhibits to the *red* query colour.

Thus, to finally rank the images, we need to take into account both the magnitude of **D** and the angle, $\angle$ **D**, that it makes with the *equidistant* line. To do this we combine the two values using a weighted sum:

$$\Re = w_1|D| + w_2\angle D, \qquad (3)$$

where lower retrieval rank values $\Re$ imply images with a closer match to all the query colours. The weights $w_1$ and $w_2$ can be adjusted to control which of the two parameters, i.e., magnitude or angle, are to dominate. We have found that values of $w_1$=0.8 and $w_2$=0.2 give the most robust results. This is to be expected since co-linearity with the equidistant line does not necessarily imply high similarity with *any* query colour. It implies that each query colour is equally close to the indexed colours. However, as $|\mathbf{D}|\rightarrow 0$, closer matches to one or more colours is implied. Thus, a greater emphasis must be placed on the magnitude component.

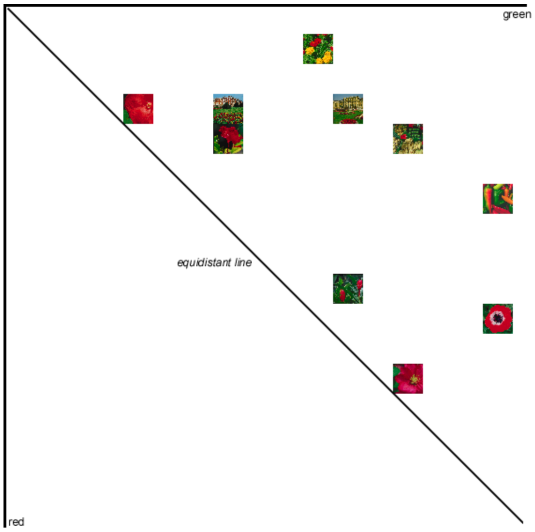


*Figure32*: A visualization of the MQDS for the two dimensional case, i.e., for two query colours. The query colours were RGB=26,153,33 (*green*) and RGB=200,7,25 (*red*). A set of retrieved images are displayed at various points in this 2-D space. Their location represents the point in space where their corresponding **D** exist.

### 3.2 Colour Exclusion

Our proposed vector approach provides a framework which easily accepts exclusion in the query process. It allows for image queries containing any number of colours to be excluded in addition to including colours in the retrieval results. As discussed above, those images which exhibit high similarity to all the query colours, are those whose corresponding MQDV **D** are collinear with the *equidistant line* and which have small magnitude. Hence, if an image contains a colour which exhibits high similarity to a query colour tagged as *exclude*, this value should affect the overall ranking by decreasing the overall similarity of the image in question; this effectively reduces the image's ranking and places it further away from the top results.

Thus, the exclusion colours similarity value should affect **D** accordingly by changing its relation to the *equidistant line* and the origin. For example, if it is found that an image contains an indexed colour which is close to an exclusion colour, the distance between the two can be used to either pull or push **D** closer or further to the ideal and accordingly affect the retrieval ranking of the given image.

To this end, we determine the similarity of each exclusion colour, (since more than 1 colour can be excluded), with the indexed representative vectors:

$$x_j = \max(S(\xi_{\mathrm{j}}, \mathrm{i}_1), \cdots, S(\xi_{\mathrm{j}}, \mathrm{i}_{\mathrm{r}})), \qquad (4)$$

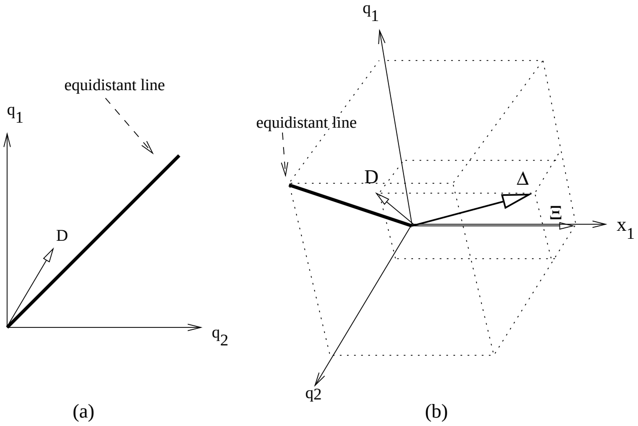
where $\xi_1, \xi_2, \ldots, \xi_m$ are the *m* colours which are to be excluded from the query results. As in (1) we build a vector comprised of the similarity values of all the exclusion colours and call this the *exclusion distance vector* (EDV), which is defined as:

$$\Xi(\xi_1, \xi_2, \cdots, \xi_m) = (x_1, x_2, \cdots, x_m)\ . \qquad (5)$$

Finally, the EDV is then merged with the MQDV to form the total query distance vector **Δ** in a new higher dimensional space, whose dimensionality is equal to *the number of query colours* + *the number of exclusion colours*:

$$\Delta = [D \quad \Xi]. \qquad (6)$$

The final retrieval rankings are then determined from the magnitude of |**Δ**| and the angle which **Δ** in (6) makes with the equidistant line of the query colour space, (i.e., the space spanned by **Δ**), without the exclusion distance vector **Ξ**. Figure 4 depicts our concept of colour exclusion. Figure 4(a) depicts a typical **D** for a 2 colour query, in its corresponding two dimensional MQDS. Figure 4(b) depicts the EDV **Ξ** and the new MQDS formed by including the similarity value of one exclusion colour. As can be seen from this figure, the inclusion of the exclusion vector **Ξ** effectively pulls **D** away from the equidistant line and at the same time increases the magnitude of the vector, forming **Δ**. Thus, the ranking of an image which contains a representative colour with high similarity to the exclusion colour will substantially increase, as compared to those images that do not contain the exclusion colour.



(a) (b)

*Figure 4*: A graphical view of the MQDS. (a) Two query colours $q_1$ and $q_2$, the MQDV **D**, and the equidistant line. (b) When an exclusion colour is specified, the vector **Ξ** pulls **D** away from the equidistant line.

# 4 Results

The concept was tested with a two colour query. More specifically the system was asked to retrieve the top 25 images from a 2000 image database which had been pre-indexed for colour as discuss in Section 2. The query asked for images which contained:

>25% *red* & >25% *green,*

where *red* and *green* have RGB values as stated in Section 3.1. The system returned images that contained colours similar to the two query colours as can be seen in Figure 5(a). For the similarity in this case, we implemented the measure in [4]. However, we are not limited to this measure and the system can use any vector similarity measure. Quantitative and qualitative analysis has shown us that the results exhibit high perceptual accuracy while also outperforming conventional colour histogram techniques [3]. More specifically, our method exhibits an average retrieval rate of 56% compared to an average retrieval rate of 35% for histogram techniques. We also tested our system for colour exclusion. Specifically, we queried the system for images which contained:

>25% *red* & >25% *green &* **exclude** *yellow,*

where *yellow* has RGB=(255,240,20). Selecting to use the same colours as tested in the two colour query above allows us to see how exclusion affects the retrieval results. Figure 5(a) depicts the query result when *red* and *green* are queried. Figure 5(b) is identical to (a) except that those images which contain *yellow* are depicted in grey-scale. These images should be removed from the top retrieval results when *yellow* is excluded in the query. Figure 5(c) shows the query results when the exclusion of *yellow* is specified. Notice how the grey-scaled images in Figure 5(b) are completely removed from the top retrieval results, and how all the retrieved images exhibit colours very similar to *red* and *green*.

## 5 Conclusion

We introduced the Multidimensional Query Distance Space, which is a query-dependent space whose dimension is defined by the number of query colours. The location of this distance vector within the space, and its relation to the origin and *equidistant line*, determines the overall retrieval ranking of a given database image. In addition, the MQDS allows for easy and effective incorporation of *colour exclusion*, i.e., where certain colours can be specified in the query to *not* be present in any of the retrieved images. By virtue of the Multidimensional Query Space, the similarity that indexed colours have to exclusion colours is used to affect the overall retrieval ranking of a given database image.

The retrieval results using this scheme prove to be very good both qualitatively and quantitatively and surpass colour histogram techniques for image retrieval.

*References:*

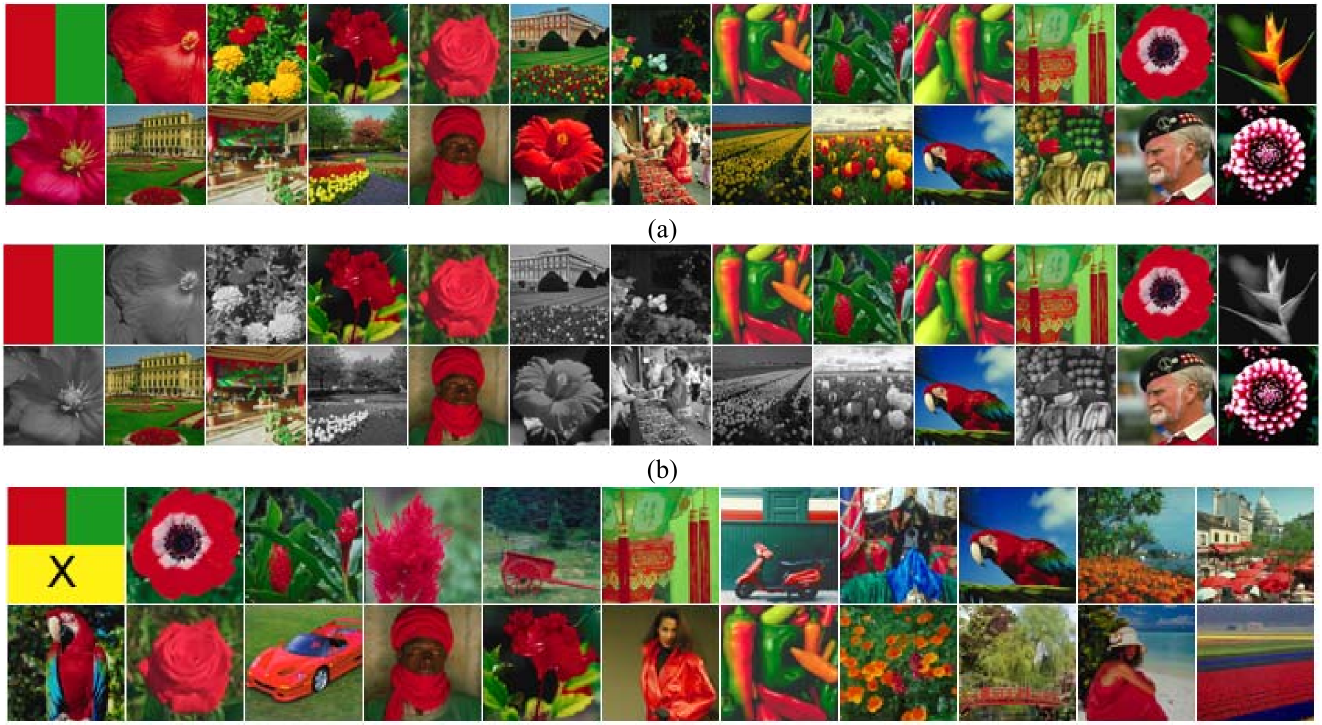
[1] Salton, G., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[2] Raghavan, V.V., Wong, S.K.M., "A critical analysis of vector space model for information retrieval", *Journal of the American Society for Information Science*, Vol.37, No.5, 1986, pp. 279-287.

[3] Androutsos, D., "Efficient Indexing and Retrieval of Colour Image Data Using a Vector-Based Technique, Ph.D. Thesis, University of Toronto, 1999.

[4] Androutsos, D., Plataniotis, K., Venetsanopoulos A.N., "A Vector Angular Distance Measure for Indexing and Retrieval of Color," *Storage & Retrieval for Image and Video Databases VII*, SPIE-3656, pp. 604-613, San Jose, USA, 1999.

[5] Swain, M.J., Ballard, D.H., "Color Indexing", *International Journal of Computer Science*, Vol. 7, No.1, 1991.



(a)

(b)

(c)

*Figure 5*: Query result for images with (a) *red* & *green* (b) Image in grey-scale are those images containing *yellow* which should not be retrieved when the query excludes *yellow* and (c) the actual results obtained from our system when querying for *red* & *green* while excluding any amount of *yellow.*