Visual Search Coursework

Andy Pack

Abstract

The efficacy of various descriptors for visual search was investigated by applying them to the MSRCv2 dataset. Descriptors of varying complexity were considered from the less discriminative global colour histogram to spatial techniques that consider both colour and shape information.

Following these results the use of principal component analysis was used to lower the dimensionality of descriptors and aid in extraction and processing. The associated Mahalanobis distance tended to provide an improvement to a standard Euclidean distance measure but the L1 generally outperformed the both.

Spatial texture techniques, specifically a combination of colour and texture, were found to be the best performing with PCA providing extra performance when applied to the correct degree. Internal parameters were generally found not to increase performance linearly but achieve a peak before degrading.

Contents

1 Introduction

An application of computer vision and visual media processing is that of visual search, the ability to quantitatively identify features of an image such that other images can be compared and ranked based on similarity.

These measured features can be arranged as a data structure or descriptor and a visual search system can be created through the extraction and comparison of these descriptors. This is an example of content based image retrieval or CBIR.

Visual search is used in consumer products to generate powerful results such as Google Lens and Google reverse image search. It also has applicability as smaller features of products such as 'related products' results.

1.1 Extraction

When arranged as three 2D arrays of intensity for each colour channel, an image can be manipulated and measured to identify features using colour and shape information. The methods for doing so have varying applicability and efficacy to a visual search system, many also have variables which can be tuned to improve performance.

1.2 Comparison

Typically a descriptor is a single column vector of numbers calculated about an image. This vector allows an image descriptor to plotted as a point in a feature space of the same dimensionality as the vector. Images that are close together in this feature space will indicate that they have similar descriptors. Methods for calculating the distance will determine how images are ranked.

2 Descriptors

2.1 Average Colour

Average colour represents one of the most basic descriptors capable of being calculated about an image, an array of three numbers for the average red, green and blue intensity values found in the image.

These three numbers represent nothing about the distribution of colour throughout the image and nothing regarding edge and shape information. The lack of either hinders it's practicality for real world applications and as a result it is not used as a descriptor on it's own in this paper. Instead average colour will be used as a sub-descriptor during spatial colour investigations aided by its low dimensionality.

2.2 Global Colour Histogram

A global colour histogram extracts colour distribution information from an image which can be used as a descriptor.

Each pixel in an image can be plotted as a point in it's 3D colour space with the axes being red, green and blue intensity values. Visually inspecting this colour space will provide information about colour scattering found throughout the image. As different resolutions of images will produce datasets of different sizes in the feature space, a descriptor must be devised that transforms this data into a resolution agnostic form which can be compared.

Each axes is partitioned into q divisions so that a histogram can be calculated for each colour channel. Each channel's intensity value, val, can be converted into an integer bin value using equation 1, where floor strips a float value into an integer by truncating all values past the decimal point.

$$bin\,val = floor\left(q \cdot \frac{val}{256}\right) \tag{1}$$

This allows each pixel to now be represented as three 'binned' values, a full RGB colour space has been reduced to three colour histograms, one for each channel. In order to arrange this as a descriptor each point should be further reduced to a single number so that a global histogram can be formed of these values. This is done by taking decimal bin integers and concatenating them into a single number in base q. For an RGB colour space, each pixel can be augmented as shown in equation 2.

$$pixel\,bin = red\,bin \cdot q^2 + green\,bin \cdot q^1 + blue\,bin \cdot q^0 \tag{2}$$

Calculating a histogram of each pixel's bin value will function as a descriptor for the image once normalised. This normalisation will remove the effect of changing resolutions of image.

Each descriptor plots an image as a point in a q^3-dimensional feature space where similarity can be computed using a suitable distance measure (L2 norm for example).

2.2.1   Efficacy

The advantage of global colour histogram over the average RGB descriptor is that amounts of colours are now represented in the descriptor. Clusters of similar colours representing objects or backgrounds will be captured and can be compared.

A global histogram, however, holds no spatial colour information, this is lost by plotting the pixels in their colour space.

This suggests that performing a pixel shuffling operation on the image will not affect the extracted descriptor which has implications on the adequacy of the methodology for a visual search system.

2.3   Spatial Colour

Spatial techniques involve descriptors that represent information from different areas of the image. This is done by dividing the image into a grid of cells and then calculating individual 'sub-descriptors' which are concatenated into a global image descriptor.

These sub-descriptors can be calculated using any appropriate method however a main consideration should be the dimensionality of the final descriptor. This can be calculated using the following equation,

$$D_{total} = W \cdot H \cdot D_{sub-descriptor}$$

Where W and H refer to the number of columns and rows of the determined grid respectively.

It would be feasible to calculate a colour histogram however this already generates a descriptor of q^3 dimensionality, where q is the number of bins.

For example using a q value of 4 and a spatial grid of 6 x 4 would produce a descriptor in 1536 dimensions, while a q of 6 with a a grid of 10 x 6 is 12,960 dimensional.

This is an extremely high value and will increase the time taken to calculate and compare descriptors.

For a spatial colour descriptor the average RGB values for each cell can be used as these sub descriptors will be three dimensional reducing the total size.

2.3.1   Efficacy

Computing a spatial colour descriptor can increase performance when highlighting the difference to a colour histogram. While a colour histogram will describe how many of each colour is present in an image, spatial colour techniques of the type described above will indicate the colours found in each area of the image. Considering an image of a cow in a field, the colour histogram will identify and count the brown pixels of the cow and the green pixels of the field, spatial colour techniques will identify an area of brown surrounded by an area of green.

2.4   Spatial Texture

Spatial texture replaces the colour sub-descriptor from before with a descriptor that reflects the texture found in the image as described by the edges that can be detected.

2.4.1   Edge Detection

Edges can be detected in an image by finding areas where neighbouring pixels have significantly different intensities.

Mathematically this can be seen as taking the first derivative of the image by convolving it with a Sobel filter. The Sobel filters are a pair of 3x3 kernels, one for each axes (see figure 1), which approximates the gradient of the grey-scale intensity of an image.

The results of convolving each filter with the image are two images that express the intensity of edges in that axes.

From here a composite edge magnitude image of the two can be calculated as shown,

$$G_{composite} = \sqrt{G_x^2 + G_y^2}$$

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad S_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$
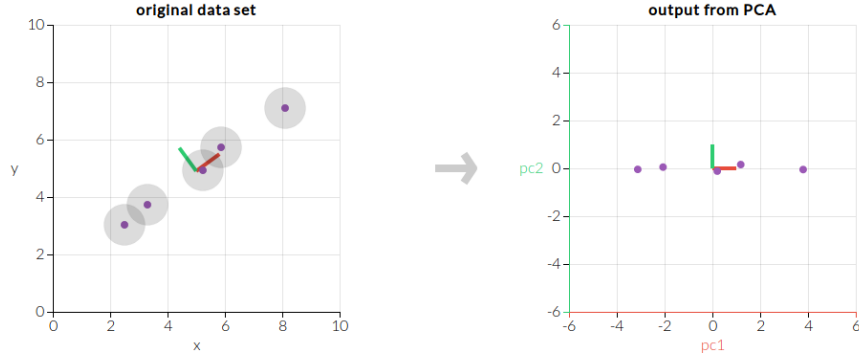
Figure 1: 3x3 Sobel filter kernels for x and y axes

Figure 2: Visual depiction of principal 2D component analysis resulting in a new reference frame[1]

With the angles of the edges calculated as follows,

$$\Theta = \arctan\left(\frac{G_y}{G_x}\right)$$

2.4.2 Application

To create a descriptor, both the angle and magnitude information will be used, the descriptor itself will reflect information about the angles of the edges found.

First the image grid cells will be thresholded using the magnitude values. Magnitude values can be seen to represent the confidence with which edges can be found and so here a decision is effectively being made as to what are and are not edges, this value can be tuned to best match the application.

Once a thresholded edge magnitude image has been found, a normalised histogram will be calculated for the angles of these edges. This histogram for each grid cell will act as the descriptor when concatenated into a vector of dimensionality, D,

$$D_{total} = W \cdot H \cdot q$$

Where q refers to the number of edge histogram bins.

3 Principal Component Analysis

When extracting a descriptor from an image an important design factor is keeping the dimensionality as low as appropriate. This is so as to reduce the effect of the 'curse of dimensionality', a set of adverse phenomena which arise when analysing high dimensional data. Each additional dimension exponentially increases the volume of the feature space, a higher dimensional descriptor requires a longer time to extract, process and store. Many of these dimensions will have little to no variation and so their addition adds complexity without adding significant extra value.

Principal component analysis is the process of identifying the orthogonal directions and magnitudes of variation in a dataset. For a multidimensional descriptor this is called the covariance. The covariance can be decomposed, or 'factorised', into matrices of eigenvectors and values defining the directions and magnitudes of the model's variations. This allows low variation dimensions to be remove by identifying vectors with low eigenvalues.

This reduced model can be used to project the dataset into the same lower dimensionality, reducing complexity but retaining the majority of variation in the remaining dimensions.
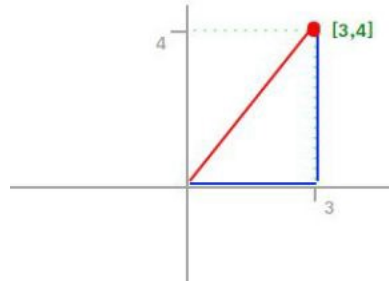
Figure 3: A single 2D Cartesian co-ordinate with it's component lengths in blue and its Euclidean magnitude in red

The process can be seen in figure 2. The left hand side shows the data presented in the root reference frame. The right hand side shows the same data plotted in the reference frame defined by it's own variation, in essence 'recentering' the axes around the data.

4   Distance Measures

Once image descriptors are plotted in a feature space, a visual search system makes comparisons by measuring the distance between them. This is calculated by finding the vector between the two points in space.

In mathematics the length of a vector is evaluated using a function referred to as a norm. Different types of norm can affect the performance of the system and therefore different norms should be used and compared.

4.1   L1 Norm

The L1 norm, or Manhattan distance, is the sum of the absolute values of the vector.

For a 2D vector, $x = (i, j)$, the L1 norm can be calculated by,

$$\|x\|_1 = |i| + |j|$$

The L1 norm of the point defined in figure 3 can be found by summing the lengths of the blue lines.

4.2   L2 Norm

The L2 norm, or Euclidean distance, is the shortest distance between two points in space, it is also referred to as the magnitude of a vector. In a three dimensional Euclidean space the magnitude of a vector, $x = (i, j, k)$, is given by,

$$\|x\|_2 = \sqrt{i^2 + j^2 + k^2}$$

It's intuitive "as the crow flies" distance measurement makes it the most commonly used norm in Euclidean space. In figure 3 the length of the red line defines this point's Euclidean distance.

4.3   Mahalanobis Distance

The Mahalanobis distance extends the L2 norm to better account for the nature of a data model defined through principal component analysis. Instead of calculating the distance between two points in a feature space, a model represents a region in space and here a more useful distance would that between a point and the closest part of that region in space.

The Mahalanobis distance allows this to be calculated by measuring the standard deviations away from the model's mean a point is, in doing so normalising for the model's shape in space. When moving from the root frame of reference to the frame defined by a model defined through PCA, the Mahalanobis distance becomes Euclidean distance. In essence we are finding the distance a point is from a model from the model's frame of reference.

The Mahalanobis distance, d, can be calculated with

$$d^2 = (x - \mu)^T C^{-1} (x - \mu)$$

7

When the covariance, C, undergoes eigendecomposition the eigenvectors, U, and eigenvalues, V, can be used to calculate the Mahalanobis distance with

$$d^2 = \left| V^{-1} U^{-1} \left(x - \mu \right) \right|$$

5   Test Methods

5.1   Dataset

For the purposes of these experiments the Microsoft MSRC[2] version 2 dataset was used. The set is made up of 591 images across 20 categories, the classifications for which can be seen in appendix A.

Worth noting about the dataset is that there are some similarities and overlap between categories which has implications on the results which can be calculated when using it.

For example category 1 is a collection of images of cows, sheep and horses on grass however cows and sheep each have their own distinct categories. Category 18 also has many similarities to category 20 with both being mainly shots of bodies of water and boats in water of varying sizes.

During the evaluation of implemented visual search techniques the classification of each image is done by referencing the group index they are named with. As such, occurrences of false negatives may increase as images that do in fact look similar as they are both, say, images of cows will be marked as not the right category and measure negatively for the performance of the method.

5.2   Precision and Recall

When comparing the effectiveness of different descriptors the main measurements are those of precision and recall.

Once the visual search system has ranked a dataset on similarity to a query image, the precision and recall can be calculated up to n images through the ranked list.

At each n the precision is defined as the number of images up to n that are classed as relevant. Higher precision values indicate better system performance and an ideal system response as n increases would be a precision of 1 until all relevant documents have been returned at which point it would gradually reduce to a minimum value of the fraction of relevant documents in the dataset. This would indicate that the system is able to select a relevant image every time one is available.

The recall is defined at n as how many of the available relevant results have been returned up to n. Higher recall values at n indicate that the system can recall relevant documents faster with less false positives and begins at 0 before increasing to a maximum of 1 as n increases when all have been returned.

While both measurements appear to reflect similar concepts there is a difference. Precision is a measure of how accurately a system can decide whether a document is relevant while recall can be thought of as a measure of a systems repeated accuracy and measures how long it takes to retrieve all relevant documents.

A system with high recall but low precision at n will indicate that the system is effectively able to retrieve all relevant documents eventually however there will be false positives within the results. Results of this quality would be advantageous when it is important to obtain all relevant results however not when the relevance of each and every one is valued.

A system with high precision but low recall at n would indicate that the system is very confident in its selection of relevant documents but may indicate an increase in false negatives where the system cannot correctly recognise a target image.

5.3   Precision Recall Curve

A way to visualise the response of a visual search system is to calculate both precision and recall for all values of n and plot each pair in a precision-recall or PR curve.

When plotted in this fashion with recall along the x axis and and precision along y the curve can be thought to plot the system performance over normalised time to retrieve the query's category set.

Perfect system performance is given by,

$$y = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & x > 1 \end{cases}$$

Figure 4: Image 10_14_s from the flower category followed by visualisation of detected edges

5.4 Methods

Results were calculated for the global colour histogram, spatial colour, spatial texture and a combined descriptor of spatial colour and texture. Principal component analysis was also conducted on the spatial colour and texture descriptor in order to identify the optimum levels of dimensionality reduction. Each descriptor has parameters that can be varied to alter it's performance. For each varying parameter a category response test was conducted.

5.4.1 Category Response

The category response test performs a number of queries on different images in order to calculate average performance values. These values are used to describe the performance of the applied descriptor, in order to make comparisons between descriptors valid the same query images are used for all investigations.

The category response aims to control for a descriptor's varying performance at each of the dataset's classifications by selecting 1 query image from each that was visually deemed to well represent the rest. The query set filenames used in these tests can be seen in appendix B. Each query iteration has precision and recall calculated for all n to allow the mean average precision to be calculated.

An average precision-recall curve was calculated by plotting the average precision and recall values for the 20 category iteration results. An example of the 20 precision-recall curves reduced to an average line is seen in appendix C.1.

Completing an iteration for each category also allows a confusion matrix to be constructed. For each iteration the top 25 results were evaluated, this number was chosen as it is similar to the average category size.

The completed confusion matrix allows the main category confusions to be identified and discussions to be made, each column represents an iteration of the category response test and the matrix has been normalised by column.

5.4.2 Global Colour Histogram

For the global colour histogram the number of bins was varied to view the effect of this on the MAP. Values between 1 and 20 were investigated.

The effect was measured using L1, Euclidean and Mahalanobis distance measures although not all bin values could be evaluated for Mahalanobis distance due to prohibitively high processing time.

5.4.3 Spatial Colour

For spatial colour, category response tests were completed as the grid dimensions were varied from 1x1 to 20x20. The resulting matrix of results were plotted as a surface to see how variations in grid size affected mean average precision.

The average precision recall curve was plotted for the best performing parameters found in these investigations. These experiments were doing using L2 (Euclidean) Distance.

5.4.4 Spatial Texture

For spatial texture the grid dimensions were brought forward from a high achieving spatial colour investigation in order to evaluate the effect of varying both bin count and threshold value. While investigating each, the other was kept the same. Following calculations for both, the best value for each was used to test the effect of varying grid dimensions. Grid dimensions were not varied as widely as for spatial colour due to the prohibitively long processing time.

The initial threshold value to be held stable during bin count investigations was picked by visualising different threshold values, examples can be seen in figures 6 through 8. A value of 0.08 was selected, the visualisation can be seen in figure 7. These experiments were doing using Euclidean Distance.

Figure 5: Image 5_15_s from the cow category followed by visualisation of detected edges
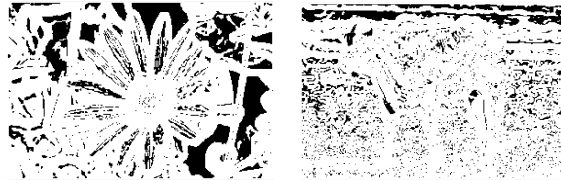

Figure 6: Angle magnitude images with threshold 0.01

Figure 7: Angle magnitude images with threshold 0.08

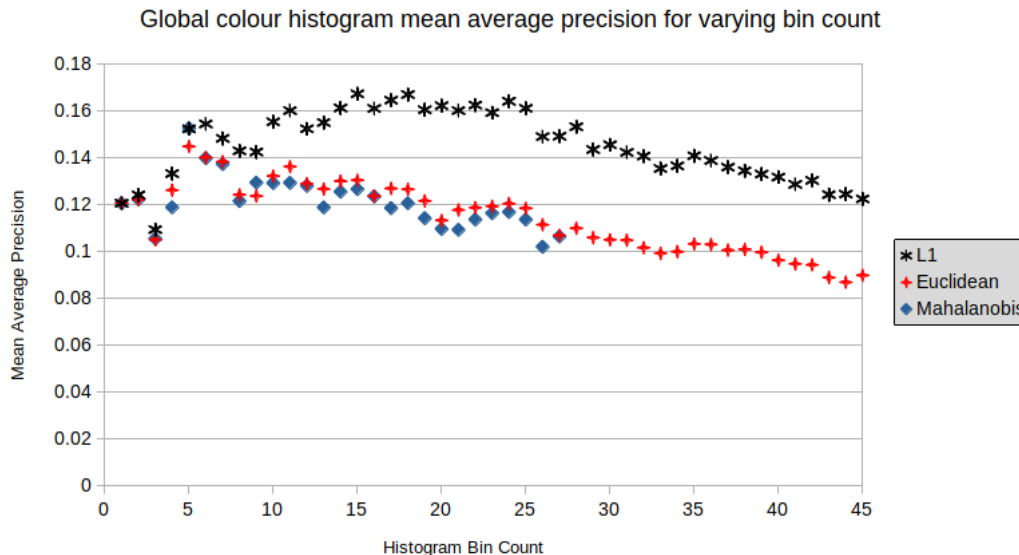Figure 8: Angle magnitude images with threshold 0.2

Figure 9: Global colour histogram mean average precision values for varying numbers of bins

5.4.5   Spatial Colour and Texture

The same spatial texture and grid dimension parameters were used when the colour and texture descriptors were combined in order to allow comparison between the results. The grid dimensions were also varied between the same ranges. These experiments were doing using Euclidean Distance.

5.4.6   Principal Component Analysis

The covariance of all descriptors is calculated in order to define a model for the dataset. In doing so dimensions with low variation can be identified and dropped by sorting the eigenvectors by magnitude of eigenvalue.

The Mahalanobis distance was found between each pair of images.

The proportion of eigenvalues that are dropped is referred to as the percentage energy reduction and this energy reduction can be varied.

To visualise the effect of varying dimensionality reduction, descriptors with previously used parameter sets were tested with varying levels of energy reduction to view it's effect on achievable mean average precision.

5.4.7   Descriptor Distance Measure

With well performing spatial parameters ascertained, mean average precision results were taken for L1, L2 (Euclidean) and Mahalanobis distance measures. These results were used to compare the performance of each distance measure with each descriptor and also as final results to summarise the relative performance of each descriptor.

6   Results

6.1   Global Colour Histogram

The effect of varying numbers of histogram bins can be seen in figure 9. L1 performed slightly better than L2 at lower bins but continues to increase in performance as Euclidean and Mahalanobis distances peak and begin to decline. Euclidean and Mahalanobis distance both peak at 5 bins (0.145 and 0.152 respectively) while the L1 norm peaks at 15 with a value of 0.167. 3 bins has a noticeably lower MAP result for all distance measures, 4 bins also seems to be slightly below the trend of increase.

The average PR curve for the best L2 mean average precision, a bin count of 5, can be seen in figure 10.

The confusion matrix for 5 bins can be seen in figure 11. The most confused category combinations were confusing cars for bikes and bikes for faces. The best classified category was book shelves.
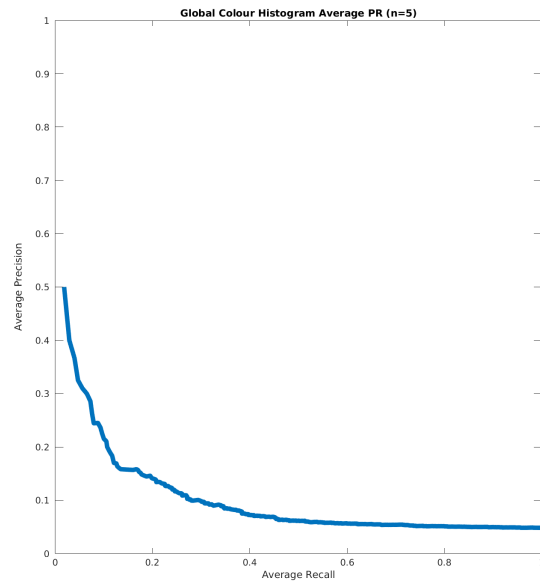
Figure 10: Global colour histogram mean precision recall curve for $n = 5$

Figure 11: Global colour histogram confusion matrix for 5 bins

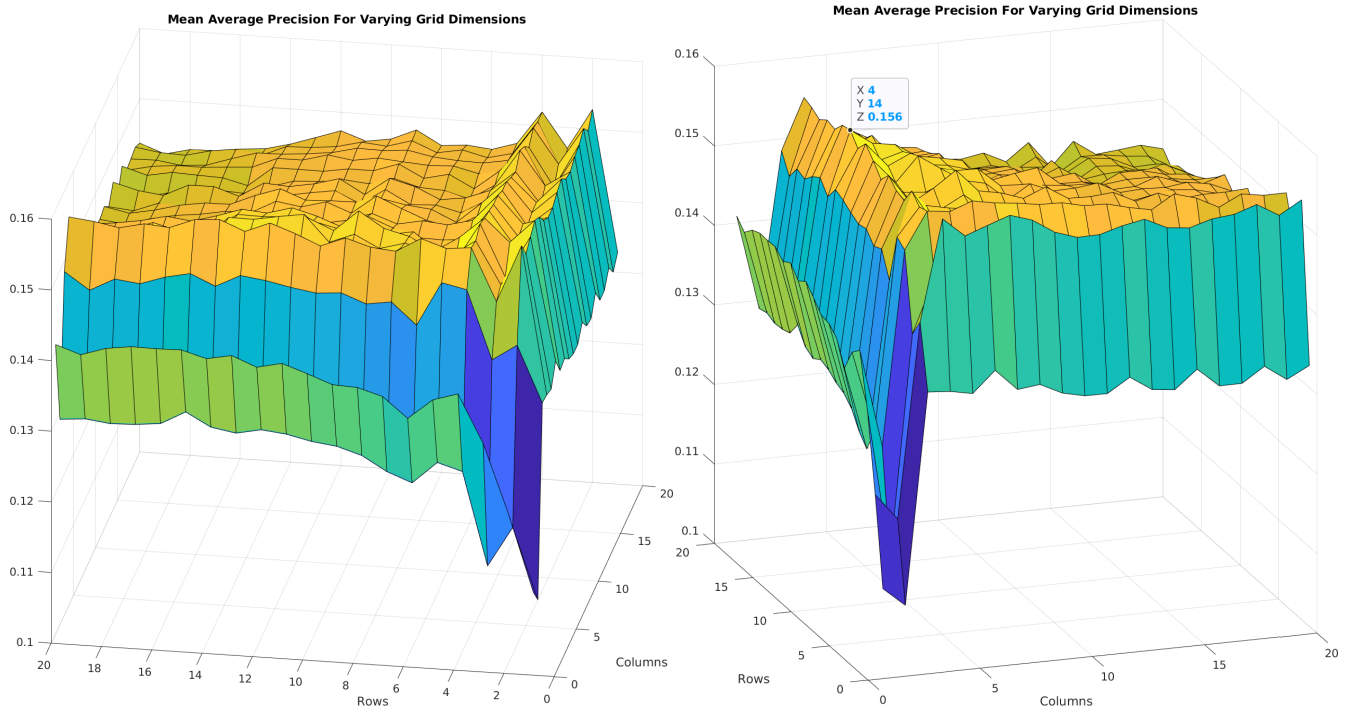| Ground Truth \ Query	Bench	Bike	Bird	Book Shelf	Building	Car	Cat	Coast	Cow	Dog	Face	Farm Animal	Flower	Human Figures	Plane	Road	Sheep	Sign	Tree	Water Features
Bench	12.0%	4.0%	8.0%		16.0%					8.0%	8.0%	16.0%		12.0%	4.0%	4.0%	4.0%	4.0%		12.0%
Bike	4.0%	36.0%			4.0%	20.0%				32.0%		8.0%	20.0%		20.0%					
Bird	4.0%	4.0%	4.0%		4.0%	4.0%				8.0%	4.0%	4.0%	8.0%	4.0%	8.0%		4.0%	4.0%		8.0%
Book Shelf	12.0%	4.0%	4.0%	48.0%		8.0%	8.0%	8.0%		24.0%				8.0%	4.0%		20.0%	4.0%	4.0%	
Building						4.0%	20.0%	28.0%	8.0%		4.0%	4.0%	4.0%		8.0%	8.0%	24.0%	8.0%	8.0%	8.0%
Car	12.0%	32.0%	4.0%		4.0%	20.0%							12.0%	16.0%	4.0%	20.0%		24.0%	16.0%	
Cat	4.0%	4.0%	12.0%	4.0%			28.0%	24.0%		4.0%				8.0%						
Coast						4.0%	8.0%	4.0%				8.0%							4.0%	4.0%
Cow	16.0%		8.0%		4.0%				8.0%			16.0%	12.0%	8.0%						8.0%
Dog	4.0%			12.0%	20.0%	4.0%	4.0%	12.0%	4.0%	12.0%	8.0%			4.0%			8.0%			8.0%
Face			4.0%	20.0%		16.0%	8.0%	8.0%		24.0%	8.0%	4.0%	20.0%				4.0%	12.0%		12.0%
Farm Animal			4.0%		8.0%				8.0%			12.0%			4.0%					4.0%
Flower	16.0%		8.0%	12.0%			8.0%		4.0%	4.0%		12.0%	12.0%	4.0%	4.0%		8.0%		4.0%	4.0%
Human Figures			4.0%		12.0%	4.0%		4.0%	16.0%	4.0%	4.0%	4.0%		12.0%	4.0%	16.0%	4.0%	4.0%	4.0%	4.0%
Plane	4.0%		4.0%		4.0%			4.0%	16.0%	4.0%	4.0%				24.0%	4.0%			4.0%	8.0%
Road	4.0%	4.0%			4.0%					4.0%	16.0%	4.0%		4.0%	4.0%	20.0%		16.0%	28.0%	
Sheep			12.0%		12.0%		4.0%		24.0%			12.0%	8.0%		4.0%		12.0%			4.0%
Sign	8.0%		8.0%	4.0%			4.0%	4.0%		4.0%					4.0%		4.0%	4.0%	4.0%	
Tree					8.0%				12.0%		8.0%				24.0%		4.0%	16.0%	28.0%	4.0%
Water Features		12.0%	16.0%			16.0%	8.0%	4.0%			4.0%	4.0%	16.0%			8.0%	4.0%			12.0%

Figure 12: Mean average precision values for varying dimensions of spatial colour grid, maximum value labelled

6.2   Spatial Colour

The surface plotted by each dimension combination's MAP can be seen in figure 12, the results up to 15x10 can be seen in appendix E.1. The MAP starts at a low value with either dimension being 1, the lowest value being for a 1x2 grid (0.1079). As rows increase from 1 to 2 a performance increase can be seen. This then decreases for 3 rows before increasing to a maxima at 4. The increase from 1 to 2 columns sees a decrease in MAP before increasing up to 4 columns.

Past 4 in either dimension the performance tends to decrease. The highest achieved MAP was for 14 rows and 4 columns with a value of 0.156. The average PR curve for this result can be seen in figure 13.

6.3   Spatial Texture

A grid of 4x4 was used initially due to it's favourable results for spatial colour. The MAP values achieved for different numbers of histogram bins can be seen in figure 14. The MAP generally increases up to 7 bins with a value of 0.210, before decreasing. Odd numbers of bins tended to achieve higher MAPs than even bins up to this maxima before linearly decreasing.

The MAP values for varying edge magnitude thresholds can be seen in figure 15. The MAP increases as threshold increases until a threshold of 0.09 (0.214). From here it decreases.

A threshold of 0.09 and bin count of 7 was used to measure the effect of grid dimensions on mean average precision, the results for which can be seen in figure 16. The highest MAP was achieved for 4 rows and 3 columns with a value of 0.231. The lowest recorded value was 0.157 for 1 row and 6 columns.

The average PR curve using the best performing of each parameter and a grid of 4x3 can be seen in figure 17.

The confusion matrix for spatial texture can be seen in figure 18. The largest confusion was the system selecting planes and classifying them as water features, 60% of the top 25 were misclassified in this fashion. The best classified categories were planes and trees.

6.4   Spatial Colour and Texture

The same parameters from spatial texture (7 bins, edge magnitude threshold 0.09) were used to investigate how grid dimensions affect mean average precision, the results for which can be seen in figure 19.
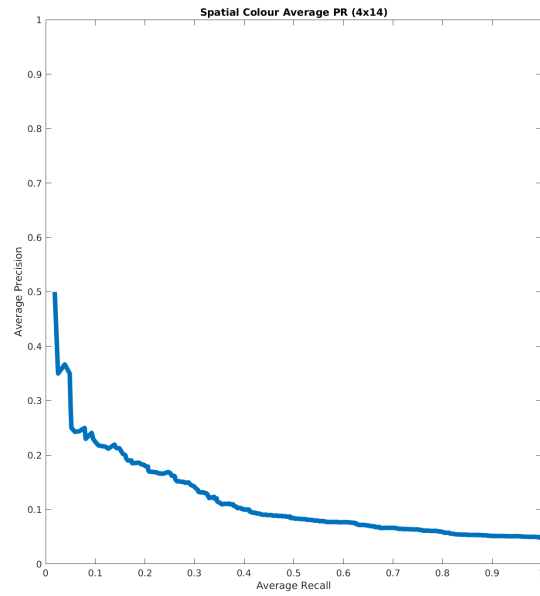
Figure 13: Mean precision recall curve for spatial colour grid 14 rows, 4 columns
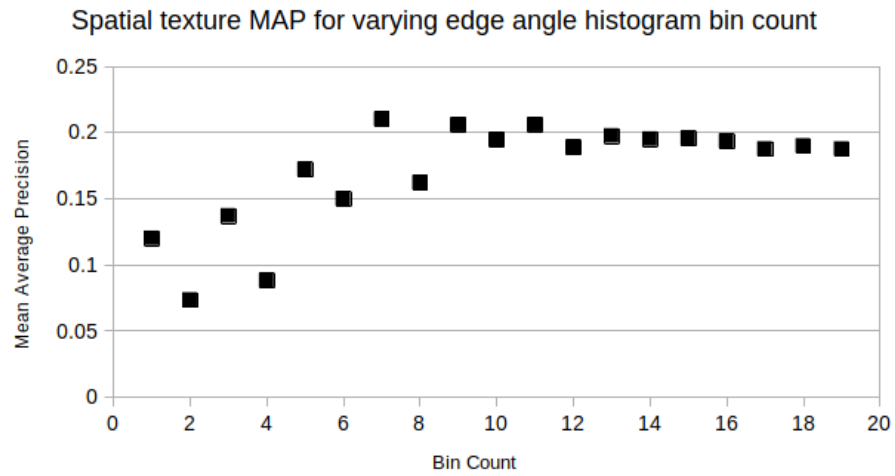
parameter

Figure 14: Spatial texture mean average precision values for varying numbers of bins, grid size 4x4, threshold 0.08
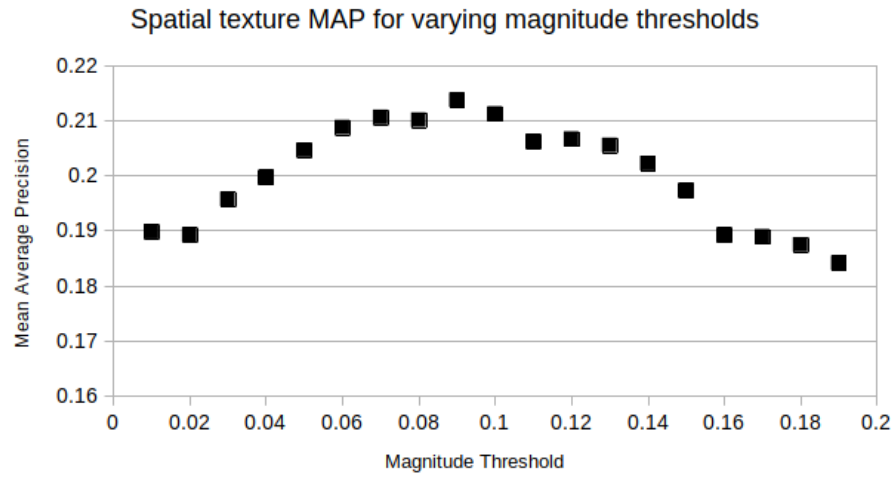
Figure 15: Spatial texture mean average precision values for varying threshold, grid size 4x4, 7 bins
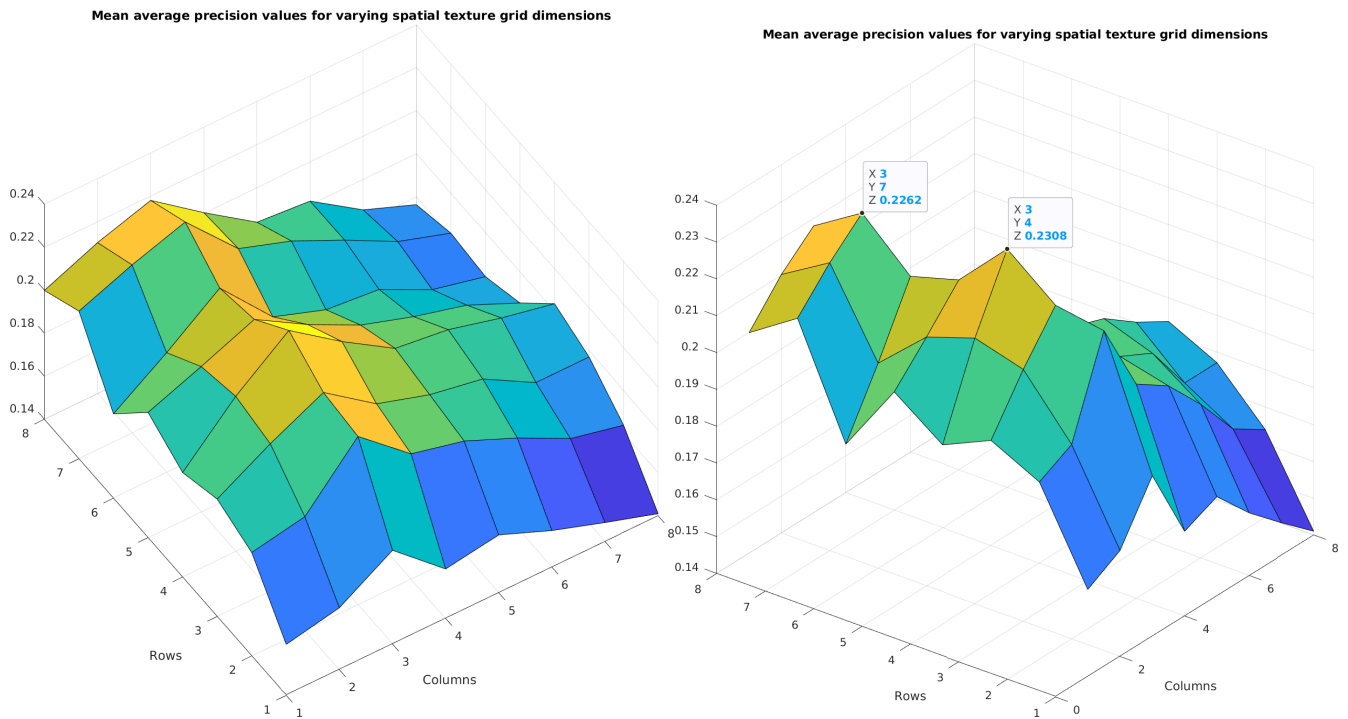

Figure 16: Mean average precision values for varying dimensions of spatial texture grid, 7 bins, threshold 0.09
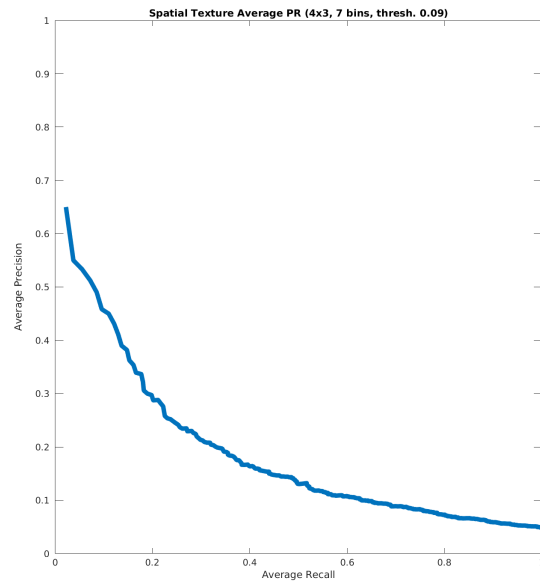
Figure 17: Mean precision recall curve for spatial texture 4x3 grid, 7 bins, threshold 0.09

Figure 18: Spatial texture confusion matrix 4x3 grid, 7 bins, threshold 0.09
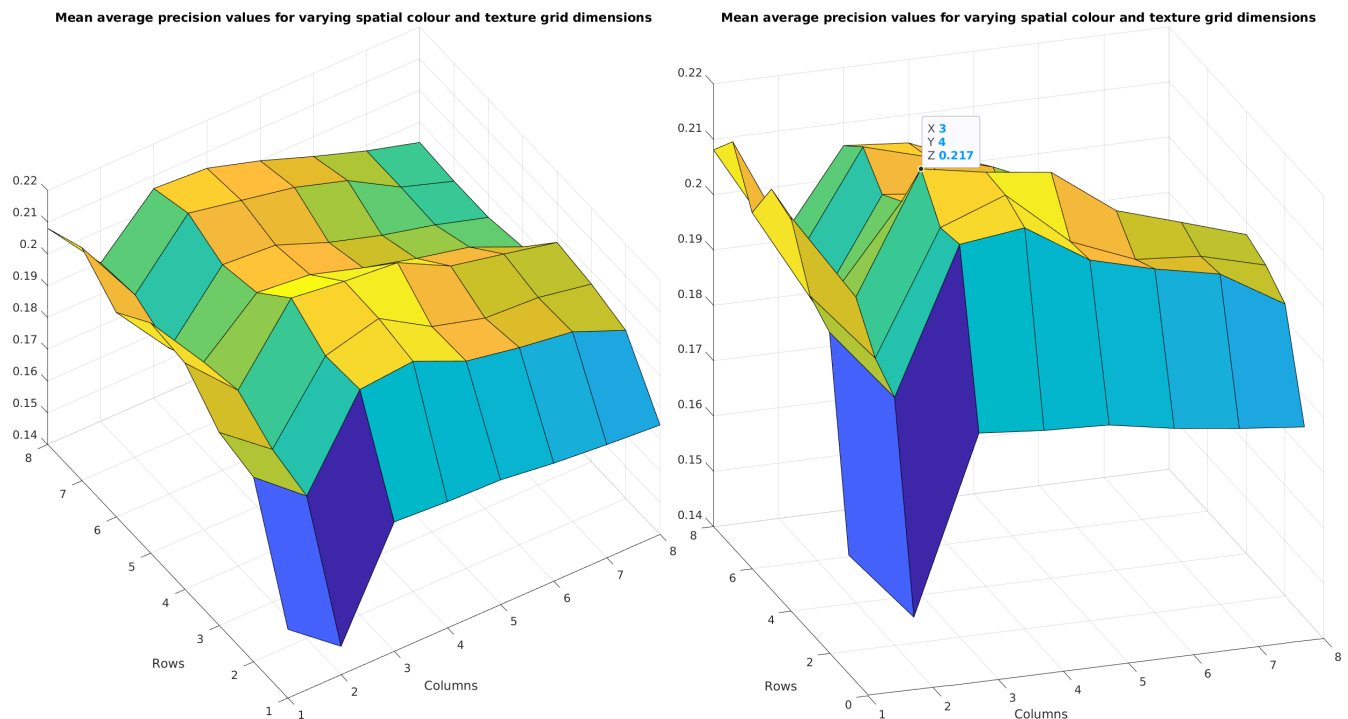
Figure 19: Mean average precision values for varying dimensions of spatial colour and texture grid, 7 bins, threshold 0.09

As rows increase the MAP increases up to 4 rows before tending to decrease past here. As columns increase from 1 to 2 the mean average precision decreases before increasing for 3 columns and then decreasing. The highest MAP was found at 4 rows and 3 columns with a value of 0.217. The lowest value was found at 2 columns and 1 row with a value of 0.1489. The average PR curve for the best performing grid dimensions can be seen in figure 20.

The confusion matrix for combined spatial colour and texture can be seen in figure 21. Here the largest errors are classifying sheep as birds, farm animals as cows and signs as book shelves. The most correctly classified images were trees and sheep.

Comparing with the spatial texture matrix, the high proportion of plane's being classed as water features decreased from 60% to 12%.

6.5 Principal Component Analysis

The spatial colour and texture parameters from before (4x3 grid, 7 bins, 0.09 threshold) were used to ascertain how varying model energy reduction affects mean average precision, the results for which can be seen in figure 22. As energy is reduced the MAP increases to a maximum between 0.5% and 4%. Beyond this reduction the mean average precision decreases to a constant value.

The red line indicates the mean average precision achieved with the same descriptor parameters without PCA.

The summary data table of these results can be seen in table 1, varying resolution's of the independent variable are in order to highlight the important values of energy reduction, namely the points at which it exceeds the non-PCA value and the highest achieved value. A higher resolution table of values can be seen in appendix F. Red values again indicate improvements over queries without PCA. The highest achieved MAP was 0.251 at 1.4% reduction. This represents a 16% increase over the 0.217 value for queries without PCA.

The confusion matrix for spatial colour and texture with PCA can be seen in figure 23. The largest confusions were classifying signs as bookshelves, cat's as the coast and sheep as birds. The most correctly classified categories were planes, book shelves and trees.

Looking to the confusion matrix without PCA, the already high confusion of sign's as bookshelves has increased from 32% to 40% despite the correct classifications also increasing from 44% to 48%.

The high confusion of sheep as bird's has decreased from 44% to 32%. However the percentage of farm animals being classed as bird's has increased from 12% to 28%.
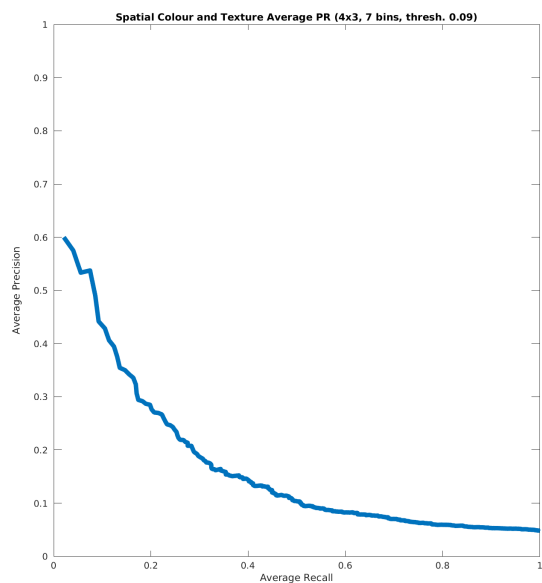
Figure 20: Mean precision recall curve for spatial colour and texture, 4 rows, 3 columns, 7 bins, threshold 0.09

Figure 21: Confusion matrix for spatial colour and texture, 4 rows, 3 columns, 7 bins, threshold 0.09

Spatial colour and texture MAP as dimensionality is reduced

Red line indicates MAP without PCA
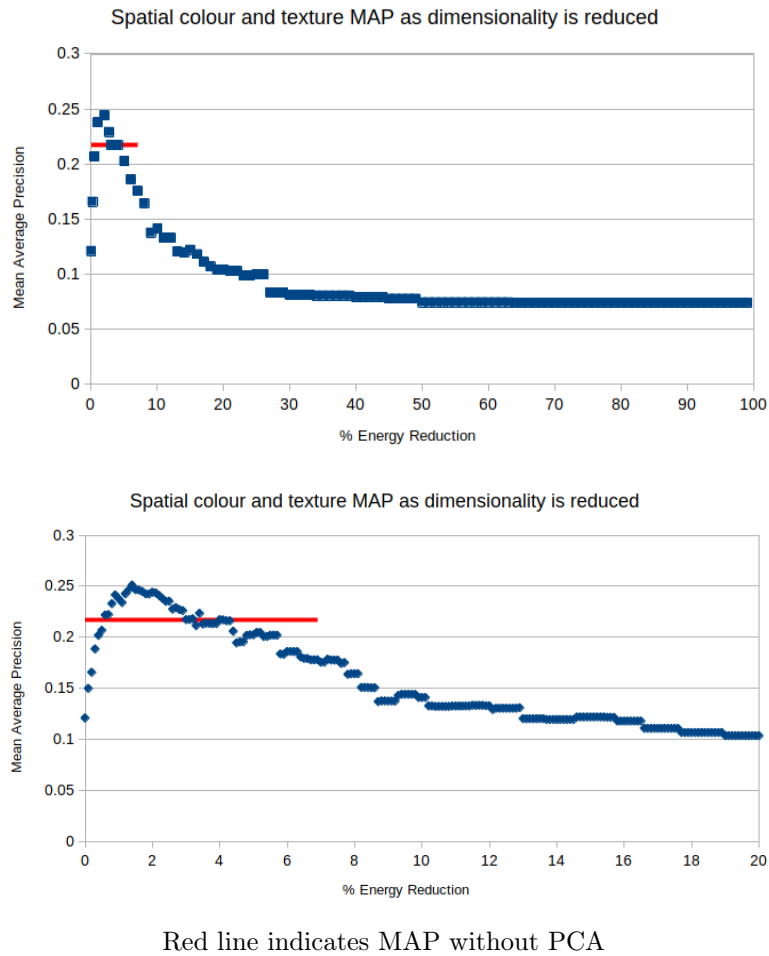
Figure 22: MAP for spatial colour and texture (4x3, 7 bins, thres. 0.09) as percentage energy is reduced

Percentage Energy Reduction	Mean Average Precision
0	0.1208
0.5	0.2067
0.6	0.2217
1	0.2377
1.4	0.251
2	0.244
3	0.2173
4	0.2172
4.1	0.217
4.2	0.2161
5	0.2023

Red values indicate improvements over query without PCA (0.217)

Table 1: Mean average precision values for spatial colour and texture (4x3, 7 bins, thres. 0.09) as dimensionality is reduced through PCA

Spatial Colour and Texture Confusion Matrix with PCA (4x3, 7 bins, thresh. 0.09)

Ground Truth \ Query	Bench	Bike	Bird	Book Shelf	Building	Car	Cat	Coast	Cow	Dog	Face	Farm Animal	Flower	Human Figures	Plane	Road	Sheep	Sign	Tree	Water Features
Bench	24.0%	4.0%			4.0%	4.0%				20.0%	8.0%		12.0%	12.0%			8.0%			4.0%
Bike	8.0%	32.0%				12.0%				12.0%	8.0%			4.0%		8.0%	4.0%	8.0%	12.0%	
Bird		12.0%	28.0%		28.0%	8.0%	24.0%	4.0%	8.0%	8.0%	4.0%	20.0%		8.0%	20.0%	8.0%		8.0%	4.0%	16.0%
Book Shelf	4.0%			48.0%										4.0%				4.0%		
Building	4.0%			12.0%	4.0%		4.0%		8.0%			4.0%		4.0%	4.0%	4.0%		24.0%	8.0%	
Car	20.0%					36.0%								4.0%		4.0%				4.0%
Cat	4.0%	4.0%			4.0%		28.0%	40.0%		8.0%	4.0%	8.0%	4.0%	4.0%						20.0%
Coast					8.0%	4.0%	4.0%	16.0%		12.0%	4.0%						8.0%			4.0%
Cow	8.0%				8.0%			4.0%	28.0%			4.0%			8.0%					
Dog		4.0%			4.0%		4.0%	8.0%	4.0%	28.0%	12.0%	4.0%	16.0%	8.0%		4.0%	8.0%	4.0%	4.0%	
Face								12.0%				4.0%		4.0%						
Farm Animal	8.0%	12.0%	28.0%		20.0%	8.0%		4.0%	24.0%		4.0%	28.0%	4.0%	4.0%			16.0%		8.0%	
Flower		4.0%	8.0%			4.0%						4.0%	24.0%		4.0%	12.0%		4.0%		
Human Figures	4.0%		4.0%					4.0%	12.0%			8.0%	12.0%	24.0%						4.0%
Plane						8.0%		16.0%		8.0%	20.0%				52.0%	4.0%			8.0%	20.0%
Road	4.0%															32.0%				
Sheep	8.0%	16.0%	32.0%		8.0%	4.0%	8.0%		8.0%		4.0%	8.0%	12.0%	12.0%		4.0%	40.0%			
Sign	4.0%	8.0%		40.0%		12.0%		8.0%		8.0%		8.0%	20.0%	8.0%			12.0%	32.0%		
Tree																			44.0%	
Water Features		4.0%			12.0%	8.0%	20.0%	16.0%		12.0%	16.0%	8.0%		4.0%	12.0%			4.0%	4.0%	32.0%

Figure 23: Spatial colour and texture confusion matrix with PCA, 4 rows, 3 columns, 7 bins, threshold 0.09

6.6   Descriptor Distance Measurements

Figure 24 shows a comparison of mean average precision values for different descriptors using their best achieved results. For spatial descriptors all have a grid of 4 by 4 in order to allow a direct comparison, spatial texture descriptors also use a bin count of 7 and edge magnitude threshold of 0.09 in order to do the same.

   The inclusion of spatial texture over solely colour derived descriptors saw a general increase in mean average precision the outlier being Mahalanobis distance for spatial texture which the lowest recorded value. L1 was the best performing distance measure overall being the best for all descriptors except spatial colour. The use of PCA and Mahalanobis distance was the second best performing distance in all cases except spatial texture. This leaves Euclidean distance performing the worst of all categories except spatial texture. The use of L1 with a spatial colour and texture descriptor had a 53% higher mean average precision than a global colour histogram with an L1 measure.

7   Discussion

7.1   Global Colour Histogram

All three of the distance measures increased in performance up to a peak value. This can be attributed to the descriptor becoming more discriminative. Low bin counts will mean that many more pixels are placed in each bin to the point that comparisons are less meaningful.

   Equally, performance decreases past a certain point as the descriptor starts to over-fit to the data. Over-fitting describes a descriptor that is too sensitive to differences in images such that it is no longer comparing meaningful variables.

   For the global colour histogram if the bin size continued to increase, each bin would eventually be comprised of only one or a handful of pixels which cannot be compared effectively. This explains why for bin count and many other descriptor parameters there is an optimum value and performance doesn't just increase linearly with that parameter.

7.2   Spatial Colour

The lower dimension configurations of the spatial colour descriptor performing worse can be explained the same way as low bin numbers for the global colour histogram. A 1x1 grid indicates that the average colour values are
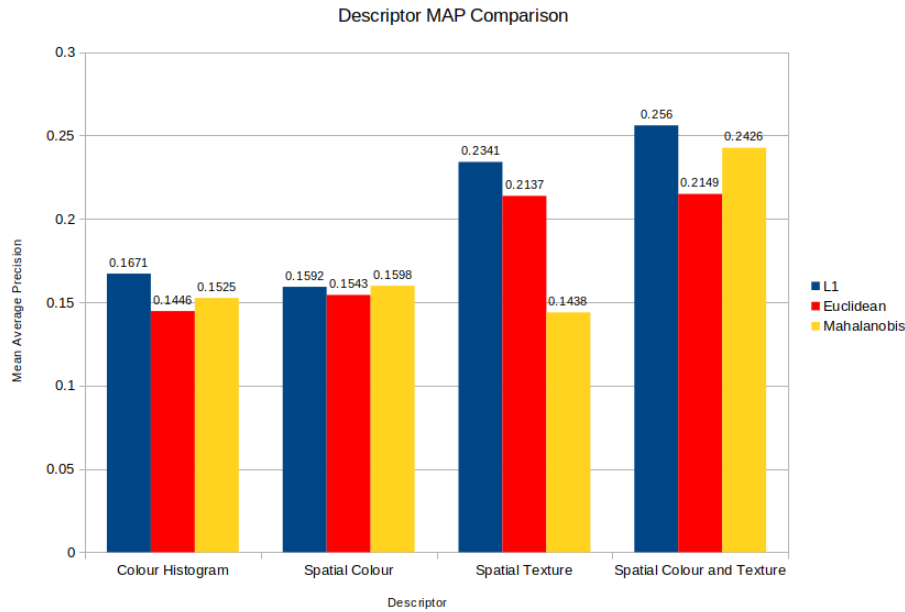
Figure 24: Mean average precision for different descriptors at same parameters

taken for the image in it's entirety, this 3D descriptor would not be particularly discriminative.

The significance of the best performing configuration having 14 rows is not particularly high. Looking to figure 12 either dimension being 4 gives similarly good results with little variation, the best being 14 by a slim margin.

An interesting result was that increasing the row count from 1 to 2 increased MAP while increasing the columns in the same fashion decreased performance.

7.3   Spatial Texture

The trend of increase in performance for increasing bin counts can be attributed to creating a more discriminative descriptor. For example, a bin count of 1 will mean that each cell's sub-descriptor is made up of only the count of pixels deemed to be edges without any further information as to which direction they are facing. As the count increases from here a maximum performance can be attained before the descriptor starts over-fitting to the data.

The interesting results of odd numbers of bins giving better than even bins could suggest that the edges of this combination of query and dataset are naturally more discriminative when quantised into odd numbers of bins as opposed to even although this would require investigation to comment further.

Worth noting is that during the spacial texture experiments optimal parameters were used as fixed variables for further testing, namely the bin count and threshold values. In doing so good results for the descriptor could be evaluated for this query however the results are query and data dependent as are the rest of the results in this coursework. Cascading optimal parameters in further experiments like this does not necessarily represent the best set of parameters for this descriptor as each optimum value was found with the others fixed.

This query and dataset's set of optimum parameters for a descriptor could be experimentally found with more complex testing methods however this is not within the scope of this work.

7.4   Spatial Colour and Texture

Looking to figure 19 this surface can be seen to be a combination of the two traced by spatial colour and texture individually. The spatial colour characteristic of decreasing performance from 1 to 2 columns but increasing from 1 to 2 rows is visible, but the maximum value at 4x3 from spatial texture is also apparent.

Looking to the confusion matrices of spatial texture and spatial colour and texture, the combinatory descriptor does not perform better in every category.

While spatial colour and texture does, for example, reduce the mis-classification of planes as water features the percentage of correct water feature classifications only increases by 4%. Spatial colour and texture performed worse at correctly classifying

- Birds

- Buildings

- Cars

- Coast

- Cows

- Human Figures

- Planes

Worth noting is the cow query classification column which shows that only 8% of the top 25 results were of the cow category. 44% were farm animals, with reference to section 5.1 this can be explained by the visual overlap between the two categories. Inspecting the dataset, half of the farm animals category are photos of groups of cows which can lead to false negatives.

Also interesting, looking to figure 24, was the slight increase in mean average precision between spatial texture and the inclusion of colour information. It could be expected that including both would provide greater improvements over applying either separately than were observed.

7.5 Principal Component Analysis

The reduction of eigenvalue energy for a model was seen to increase achievable MAP for the spatial colour and texture descriptor when done within a specific range. The initial increase in performance can be attributed to removing dimensions of low variance and importance. In doing so the remaining descriptor can be more discriminative.

However past this point relevant information starts being removed resulting in degradation of performance.

In summary the use of PCA and Mahalanobis distance does not inherently increase performance but using a reasonable level of energy reduction such as 2% can increase performance.

7.6 Comparisons

Looking to figure 24 it can be seen that spatial texture descriptors can achieve better mean average precision values than solely colour information based descriptors. This indicates that using spatial information can form a more discriminative descriptor than without.

The slight performance gain from the colour histogram to spatial colour suggests that when looking solely at colour, spatial information does not necessarily well surpass a colour histogram.

When thinking about a picture of a cow like in figure 5, the spatial colour descriptor will identify an average of green surrounding an average of brown. The colour histogram however still identifies both the brown and green pixels and encodes them.

While it may have been thought that spatial techniques were better as a whole than the global colour histogram, they in fact perform comparably indicating that they may have different applications.

Spatial texture descriptors represent the largest increase in performance. This can be attributed to spatial texture techniques attempting to identify and describe the shapes of objects within the image as opposed to just the colours present.

Colour exclusive methods can be less discriminative as different objects can generate similar descriptors by being similar colours. This could explain the global colour histogram frequently confusing cars and bikes as both have similar metallic colours. Using shape information allows for discrimination between images such as those seen in figure 25 where both are white objects on a green background. Here using shape information provides an opportunity to distinguish between the two.

The act of reducing the dimensionality of most of the descriptors improved performance however this only occurs when the optimum level of reduction is achieved. Looking again to figure 22 only the maxima of the curve represents results that were better than could be achieved without reduction. Without identifying this range the performance can be worse, 2% appears to be a reasonable range to reduce by. When using this value for all descriptors a performance increase was generally observed indicating that it may be worth considering in most instances.

Figure 25: Example of images with comparable colour information but different shapes

8   Conclusions

To summarise, each descriptor has varying efficacy for different applications and the context should be identified before selecting one over the other.

Optimum internal parameters for each are data and query dependedent but well performing values can be obtained experimentally.

Spatial texture techniques were found to perform better than exclusively colour based methods such as a global colour histogram. While the combination of both colour and texture information didn't provide as much of an improvement over either separately as expected, the combination of both colour and texture information as part of the descriptor was the best performing while using the L1 norm.

The use of appropriate amounts (approximately 2%) of dimensionality reduction following principal component analysis provided performance increases over the standard L2 norm, however following investigations into difference distance measures the L1 norm tended to be best performing method.

References

[1] V. Powell and L. Lehe, *Principal component analysis*. [Online]. Available: `http://setosa.io/ev/principal-component-analysis/`.

[2] *Image understanding*, Jan. 2000. [Online]. Available: `https://www.microsoft.com/en-us/research/project/image-understanding/`.

A   MSRCv2 Dataset Classifications

Category Index	Category Classification
1	Farm Animal
2	Tree
3	Building
4	Plane
5	Cow
6	Face
7	Car
8	Bike
9	Sheep
10	Flower
11	Sign
12	Bird
13	Books
14	Bench
15	Cat
16	Dog
17	Road
18	Water Features
19	Human Figures
20	Coast

B   Query Set

Category Index	Query Filename
1	1_10_s.bmp
2	2_16_s.bmp
3	3_12_s.bmp
4	4_4_s.bmp
5	5_15_s.bmp
6	6_14_s.bmp
7	7_17_s.bmp
8	8_15_s.bmp
9	9_1_s.bmp
10	10_14_s.bmp
11	11_8_s.bmp
12	12_26_s.bmp
13	13_10_s.bmp
14	14_10_s.bmp
15	15_8_s.bmp
16	16_10_s.bmp
17	17_16_s.bmp
18	18_5_s.bmp
19	19_15_s.bmp
20	20_12_s.bmp

C   PR Curves

C.1   Global Colour Histogram, $n = 4$
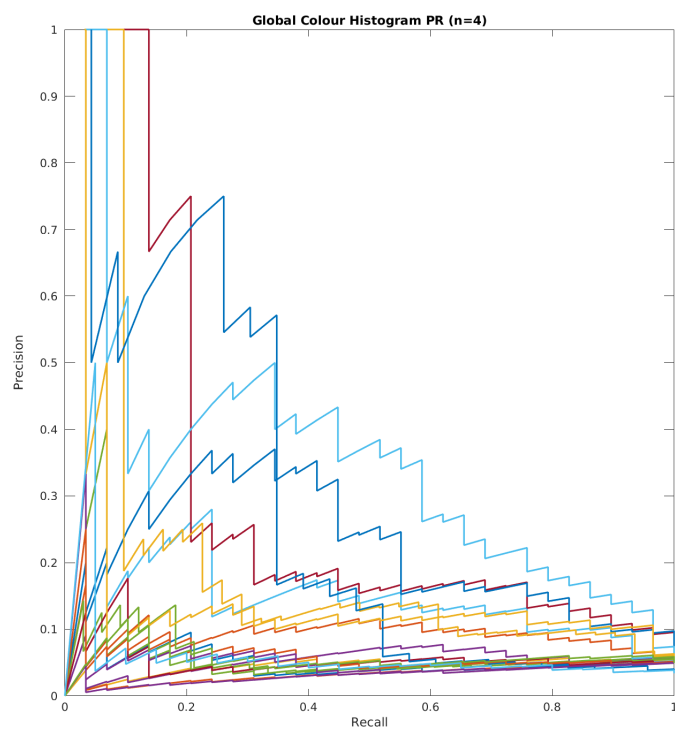
D   Global Colour Histogram Bin Count MAP Results

Number of bins	Mean Average Precision	Average Precision Standard Deviation
1	0.1204	0.2118
2	0.126	0.1002
3	0.1049	0.0678
4	0.126	0.1002
5	0.1446	0.1034
6	0.14	0.1209
7	0.1381	0.1193
8	0.124	0.1127
9	0.1235	0.1166
10	0.132	0.1241
11	0.136	0.122
12	0.1288	0.1171
13	0.1265	0.1161
14	0.1299	0.1136
15	0.1303	0.1105
16	0.1234	0.117
18	0.1264	0.1148
20	0.1131	0.107
22	0.1186	0.1158
24	0.1203	0.1135
26	0.1113	0.0981
28	0.1098	0.0963
30	0.1049	0.0969
32	0.1015	0.0914
34	0.0998	0.0932
36	0.1028	0.0924
38	0.1008	0.0884

E   Spatial Mean Average Precision Tables

E.1   Colour

Grid Size	1	2	3	4	5	6	7	8	9	10
1	0.1103	0.1079	0.1345	0.1342	0.1337	0.1363	0.1336	0.1339	0.1326	0.1319
2	0.1213	0.1179	0.1424	0.1551	0.1527	0.1537	0.1547	0.1538	0.1519	0.151
3	0.131	0.1123	0.1401	0.1469	0.148	0.1477	0.1513	0.1495	0.1493	0.1499
4	0.1377	0.1254	0.1498	0.1543	0.154	0.155	0.1545	0.1542	0.1537	0.1536
5	0.1366	0.1264	0.1505	0.154	0.1516	0.1494	0.1515	0.1497	0.1485	0.1499
6	0.1338	0.1233	0.1443	0.1552	0.1507	0.1509	0.1493	0.1478	0.1482	0.149
7	0.1366	0.1246	0.1473	0.1531	0.1522	0.1516	0.1512	0.1505	0.1483	0.1493
8	0.1377	0.1267	0.1468	0.1535	0.1498	0.1506	0.1521	0.1499	0.1494	0.1486
9	0.1378	0.1278	0.1481	0.1546	0.1507	0.1524	0.1543	0.1514	0.1509	0.1499
10	0.1392	0.1282	0.1479	0.1532	0.1513	0.151	0.1513	0.1502	0.1491	0.1487
11	0.1404	0.129	0.1484	0.1547	0.1516	0.1547	0.1528	0.151	0.1504	0.148
12	0.1396	0.1294	0.1492	0.1559	0.1529	0.1542	0.1522	0.1514	0.1513	0.1489
13	0.141	0.1287	0.1496	0.1545	0.1527	0.1535	0.1516	0.1512	0.1504	0.1488
14	0.1403	0.1293	0.1479	0.156	0.154	0.1536	0.1518	0.1513	0.1507	0.1477
15	0.1413	0.1312	0.1494	0.1557	0.1515	0.1503	0.15	0.1489	0.1495	0.1464

E.2   Texture

Grid Size	1	2	3	4	5	6	7	8
1	0.1637	0.1688	0.1836	0.163	0.1669	0.157	0.1488	0.141
2	0.1881	0.1928	0.2182	0.1982	0.1922	0.1817	0.1695	0.1638
3	0.1946	0.2084	0.2203	0.2031	0.1954	0.1908	0.1772	0.1772
4	0.1886	0.2121	0.2308	0.2137	0.1986	0.1954	0.1889	0.1836
5	0.1982	0.2076	0.2176	0.2033	0.1911	0.1831	0.1765	0.1656
6	0.1792	0.1957	0.2138	0.1885	0.1799	0.1774	0.1654	0.1595
7	0.2086	0.2182	0.2262	0.2019	0.1935	0.182	0.1694	0.1614
8	0.1999	0.2102	0.2179	0.2003	0.1839	0.1818	0.1658	0.1563

E.3   Colour and Texture

Grid Size	1	2	3	4	5	6	7	8
1	0.1616	0.1489	0.1807	0.1797	0.1792	0.1771	0.1756	0.1744
2	0.1981	0.1848	0.211	0.2125	0.2052	0.2021	0.1997	0.1928
3	0.2008	0.1881	0.2102	0.2146	0.2047	0.2	0.1991	0.196
4	0.2113	0.1953	0.217	0.2149	0.2134	0.205	0.2014	0.1977
5	0.2125	0.1928	0.2073	0.2056	0.2004	0.196	0.1922	0.1849
6	0.2044	0.1857	0.2047	0.2035	0.197	0.1919	0.1878	0.1828
7	0.2134	0.1912	0.2095	0.2062	0.2029	0.1977	0.1883	0.1828
8	0.2081	0.1898	0.2059	0.2049	0.1999	0.1939	0.1883	0.1836

F   Principal Component Analysis MAP values

Percentage Energy Reduction	Mean Average Precision
0	0.1208
0.1	0.1498
0.2	0.1656
0.3	0.1885
0.4	0.2017
0.5	0.2067
0.6	0.2217
0.7	0.2222
0.8	0.2329
0.9	0.2414
1	0.2377
1.1	0.234
1.2	0.2425
1.3	0.2462
1.4	0.251
1.5	0.2465
1.6	0.2463
1.7	0.2448
1.8	0.2425
1.9	0.2423
2	0.244
2.1	0.2434
2.2	0.2409
2.3	0.2379
2.4	0.2352
2.5	0.2353
2.6	0.2274
2.7	0.229
2.8	0.2271
2.9	0.2261
3	0.2173
3.1	0.2172
3.2	0.218
3.3	0.2114
3.4	0.2234
3.5	0.2129
3.6	0.2137
3.7	0.2134
3.8	0.2132
3.9	0.2133
4	0.2172
4.1	0.217
4.2	0.2161
4.3	0.2161
4.4	0.2058
4.5	0.1944