Visual Search Coursework

Andy Pack (6420013)

Abstract

abstract

Contents

1 Introduction

An application of computer vision and visual media processing is that of viusal 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 composed of the extraction and comparison of these descriptors. It is an example of content based image retrieval or CBIR.

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.

1.3 Applications

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.

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 hold no information about the distribution of colour throughout the image and no information based on edge and shape information. The lack of either hinders it's applicability to any real world problems. The only advantage would be the speed of calculation.

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 for each pixel. 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 a 3D point of three 'binned' values, a full RGB colour space has been reduced to three colour histrograms, 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 by count. 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 similiarity can be computed using a suitable distance measure (L1 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 calculating descriptors tht are discriminative between colour and shape information in different regions 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 the global image descriptor.

These sub-descriptors can be calculated using any approprate method however a main consideration should be the dimensionality of the final descriptor. This can be calculted 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 desciptor of q^3 dimensionality, where q is the number of divisions.

For example using a q value of 4 and a spatial grid of 6 x 4 would produce a descriptor in 1536 dimenions, 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 value.

2.3.1 Efficacy

Computing a spatial 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 in the middle of an image surrounded by an area of green.

2.4 Spatial Texture

Spatial texture replaces the colour sub-desciptor 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 greyscale intensity of an image.

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad S_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Figure 1: 3x3 Sobel filter kernels for x and y axes

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,

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

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 angle 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 applcation.

Once a thresholded edge maginute image has been found, a normalised histogram will be calculated for the angles of these edges. This histograms of 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 Distance Measures

3.1 L1 Norm

4 Test Methods

4.1 Dataset

For the purposes of these experiments the Microsoft MSRC[1] 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.

4.2 Precision and Recall

When comapring 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 accuracy 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 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 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 would indicate that the system is able to very confident in its selection of relevant documents but may indicate an increase in false negatives.

4.3 Precision Recall Curve

A way to visualise the response of a system is to calculate both precision and recall at each n and plot both as what is known as a precision-recall curve or PR curve.

4.4 Methods

5 Results

6 Discussion

7 Conclusions

References

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

A MSRC 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	Book Shelf
14	Bench
15	Cat
16	Dog
17	Road
18	Water Features
19	Human Figures
20	Coast