INVESTIGATION OF A KOLMOGOROV COMPLEXITY BASED SIMILARITY METRIC FOR CONTENT BASED IMAGE RETRIEVAL

By

SUPAT SUPAMAHITORN

Bachelor of Engineering

Chulalongkorn University

Bangkok, Thailand

2001

Summited to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2004

INVESTIGATION OF A KOLMOGOROV COMPLEXITY BASED SIMILARITY METRIC FOR CONTENT BASED IMAGE RETRIEVAL

Thesis Approved:

Dr. Douglas R. Heisterkamp
Thesis Advisor

Dr. John P. Chandler

Dr. H. K. Dai

Dr. Gordon Emsile
Dean of Graduate College

ACKNOWLEDGMENT

I would like to thank Dr. Douglas R. Heisterkamp, Dr. John P. Chandler, and Dr. H. K. Dai for giving advises and comments on this paper as well as my family who support me in everything.

TABLE OF CONTENT

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

LIST OF SYMBOLS

$D(x,y)$	Information distance between x and y
$d(x,y)$	Normalized information distance between x and y
$d_s(x,y)$	Normalized information distance between x and y
$K(x)$	Kolmogorov complexity of x
$K(x\|y)$	Conditional Kolmogorov complexity of x relative to y
$K(x,y)$	Kolmogorov complexity of x and y
$K(xy)$	Kolmogorov complexity of x concatenated with y
x^*	The shortest binary program for x
z^2	The approximated chi square value

1. Introduction

The amounts of information in many fields, especially in the computer field, are increasing dramatically. Therefore, retrieving a particular piece of information from all of the information is often a problem. However, there are quite a number of retrieving methods that yield reliable and effective results, for example, website search engines and text searching algorithms using string matching methods. The text information has a small number of characters which would be easier to search, comparing to the other kind of information which have more variability and complexity, such as, audio, image, and video. This paper concerns about the image information that have spatial content and more complexity. Some image search engines retrieve the images by using the image filenames matching that might not be the best method, because filenames and image information are not totally dependent. The ideal image search engine would be the one that its input is an image and the output is a set of images which are closely similar to the input image in the image information perspective. Therefore, we need to use the image contents in the image comparison algorithm. One of the comparison results is the information distance between the image. Since the attributes of the images can be different in size, dimension and color, we normalize the information distance. By using normalized information distance, we can compare the information distance of a pair of images to another pair of images.

The normalized information distance that we are using is based on Kolmogorov complexity. In other words, we use an approximated Kolmogorov complexity based similarity metric in the image retrieval, because it yields good results in many previous works.

The Kolmogorov complexity is uncomputable. Therefore, the approximated measurement is needed. In the experiment, we use a number of methods to compute the approximated Kolmogorov complexity, in order to achieve the best result. Furthermore, we compare the results with the random method, retrieving the images in random manner. As a result, we can show whether the proposed methods and random method are statistically different. Also, we compare the Kolmogorov complexity approximation methods if they are statistically different.

The detailed contents of this thesis are arranged as the following chapters. Chapter 1 is the introduction. Chapter 2 , Preliminaries, shows the ideas and theories. Chapter 3 reviews some of the selected previous works. Chapter 4 is about proposed approach and the hypothesis. Chapter 5 explains how to set up the experiment to solve the problem. Chapter 6 shows the experimental results. Chapter 7 gives the conclusions.

2. Preliminaries

This section explores the idea of the information distance in general and the comparable information distance, that is the normalized information distance, which based on Kolmogorov complexity, as well as the idea of using compressors to compute the approximated Kolmogorov complexity. Most of the usage of the ideas and theories in the selected previous works is concerned in the next section.

2.1 Information distance

The information distance is basically the difference in content between two data. In [20], information distance is a distance function D with nonnegative real values, defined on the Cartesian product X x X of a set X is called a metric on X if for every x,y,z are members of X:

- D(x,y) = 0 if x = y (the identity axiom);
- D(x,y) + D(y,z) >= D(x,z) (the triangle inequality);
- D(x,y) = D(y,x) (the symmetry axiom).

Therefore, the idea is to use the information distance as a measurement for comparing images.

2.2 Kolmogorov complexity

Kolmogorov complexity or algorithmic entropy [17, 20], K(x) of a string x is the length of the shortest binary program to compute x on an appropriate universal computer. The conditional Kolmogorov complexity K(x|y) of x relative to y is defined similarly as the length of a shortest program to compute x if y is an auxiliary input to the computation. x* denotes a shortest program for x, hence |x*| = K(x). We use the notation K(x,y) for the length of a shortest binary program that prints out x, y, and a description how to tell them apart. The information distance, [14], is the length of a shortest binary program that generates x from y as and also generates y from x, equals

$$D(x,y) = \mathrm{Max}\{\ K(y|x)\ ,\ K(x|y)\ \} \tag{2.1}$$

2.3 Normalized information distances

In [14], they defined the general information distance, the similarity metric, for text information.

There are two definitions of the normalized information distance. The first normalized information distance is defined as

$$d_s(x,y) = \frac{K(x|y^*)+K(y|x^*)}{K(x,y)} \tag{2.2}$$

A deep theorem in Kolmogorov complexity states that K(x)-K(x|y) = K(y)-K(y|x), approximately [2]. That is, the amount of information x knows about y is the same as the amount of information y knows about x. Therefore,

$$K(x|y) = K(x,y) - K(y) \tag{2.3}$$

The first normalized information distance can be defined as

$$d_s(x,y) = \frac{1 - (K(x) - K(x|y^*))}{K(x,y)} \tag{2.4}$$

The second normalized information distance can be defined as

$$d(x,y) = \frac{Max\{\ K(x|y^*)\ ,\ K(y|x^*)\ \}}{Max\{\ K(x)\ ,\ K(y)\ \}} \tag{2.5}$$

The more details of the theory of Kolmogorov complexity can be found in [14, 20].

2.4 Kolmogorov complexity approximation method

As we know, the compression algorithms reduce the size of the data by compress them, for example, Huffman's code algorithm compresses the data by encoding the repeated data to make the shorter data representation. By using the compression algorithms, we can use the compressed size of data as an approximation of the length of the shortest binary program to produce that data, which is Kolmogorov complexity of that data, K(x), where x represents the data.

In addition, the idea of compressing the repeated data in the compression algorithms could be used to approximate the Kolmogorov of x and y, K(x,y), where x and y are the data, by combining the data together and compress them. The repeated contents of the both data will be compressed together and the size of the combined data will be minimized.

3. Related works

There are many related works about using Kolmogorov complexity and information distance. Some of them used the Kolmogorov complexity in different ways, some of them used it in the normalized information distance to compare some data, most of them are text data. The following reviewed selected works will explore the idea of using the theories and ideas in the preliminary section that has been addressed before.

3.1 Image processing

The unique characteristic complexity of the images could be used in the image processing. In [20] and [22], they used Kolmogorov complexity to locate text in images. They claimed that the images that are used to hide the texts inside are usually simple. That is, the background of the text hidden images have less amount of Kolmogorov complexity than the others. The experimental results have shown that for such simple images, adding a text increases their complexity. Therefore, if an image is mostly simple, but there is one area that has more complexity than the others, then it is a good indication of the area that may contain some texts.

The next previous works will show more about Kolmogorov complexity application in the data comparison.

3.2 Source code plagiarism detection

Program plagiarism detection system in [4] uses an information based sequence distance which based on Kolmogorov complexity. Also, they measure the shared parts of the programs by using compression algorithm. The compression algorithm that they developed also can handle approximate matches by searching for approximately duplicated substrings and encodes mismatches if they provide benefits to the compression ratio.

3.3 Hierarchical clustering for language and evolutionary tree

The evolutionary tree and the language classification tree experiments in [14] also used normalized information distance to built them. In the evolutionary tree experiment, they developed new measures, combined k-mer approach with (2.2) and (2.5). Consider the length-k substrings of the DNA sequence as words of the sequence. They denoted the number of distinct, possibly overlapping, k-length words in a sequence x by N(x) and then, with k large enough, they use N(x) and N(x|y) as a rough approximation to K(x) and K(x|y), respectively, where N(x|y) is defined as N(xy) – N(y).

4. Proposed approach and hypothesis questions

The proposed approach is to use an approximated Kolmogorov complexity based similarity metric in the image retrieval problem. As we know, the Kolmogorov complexity is uncomputable, therefore, in order to achieve the closest approximation, we propose a number of methods to use in the experiment. The key of the approximation methods is to use compression methods. However, the efficiency of the selected compressor is important because the better compressor would provide the better results, which is closed to the theoretical results. In addition, the different image concatenation methods change the information in the spatial perspective that would yield the different results. Also, we want to have some image concatenation methods that maintain the aspect ratio of the images because it is an attribute of the image.

Afterwards, we could derive some hypothesis questions that need to be answered in the experiment as follow.

- Does the image retrieval using approximated normalized information distance method generate classification results that are statistically different from the random method?

In this hypothesis question, we seek to determine if the proposed methods are actually capturing some information.

- Do the different combination ofimage concatenation and compression methods provide the statistically significant difference in results?
- Do the different normalized information distances provide the statistically significant difference in results?

The last two questions are to show whether the combination of the Kolmogorov complexity approximation methods yield any differences in the experimental results.

5. Experiment methodology

This section shows the experiment setups using the preliminary ideas and theories in the proposed approach. Since the compressors will be used to estimate the Kolmogorov complexity, the proper list of compressors will be selected to create the reasonable results. Also, we want quite a number of various types of concatenation method to show the results form many setups. Not only the testing method setup is important, but also, the image data set selection is crucial, because good image data set will yield pertinent result. Afterwards, the hypothesis questions will be proved whether accepted or rejected by using some statistical methods on the experimental results.

5.1 Compressors selection

There are so many kinds of compressor that are available. Anyway, we decided to use some popular compressors that are used in many kinds of information and some compressors that especially designed to use on images. However, we do not want to lose any contents of the images during the compressing process which means all the compression algorithms that will be using are loseless. Therefore, we decided to use the following compressors.

1) gzip compressor [8]

The gzip compression algorithm finds repeated strings in the input data with 32K bytes window size and compress them. Since the images can be scanned their content in many different ways, we decided to scan the images in two ways, horizontal and vertical scans. Both scan methods start from the top-left corner of the image. In horizontal scan case, the scan goes from left to right to the end of the row and then process the next row until the end of image. In vertical scan case, the scan goes from top to bottom to the end of the column and then process the next column until the end of image.

2) JPEG2000 compressor [13]

JPEG2000 is the latest series of standards from the JPEG committee. The JPEG2000 uses wavelet technology based compression techniques and it can allow an image to be retained without any distortion or loss, unlike the lossy version, present JPEG. An example of wavelet compression is shown in Figure 1.

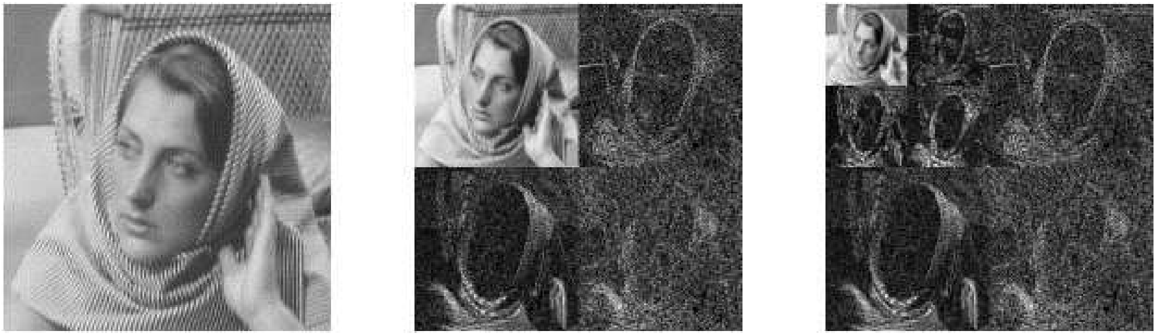

Figure 1. Wavelet compression example

The more information about JPEG2000 can be found in [13].

5.2 Image concatenations selection

The way to combine the two images together is a part of the Kolmogorov complexity approximation. Therefore, it is important to explore as many as concatenation methods. Also, the combinations of the concatenation methods and compressors yield different results. Some compressors have their limits, for example, the gzip compressor has 32K bytes window size, which is small relative to the image size. As a result, the gzip compressor need some image concatenation methods to compensate the limitation.

We designed a number of the image concatenation methods which would explore the ideas and search for the best compression results. The concatenation methods can be categorized into the following items.

1) Raw image concatenation

The Raw concatenation is to view the images as one dimensional data and append them together. This method is the simplest way to do and it does not require any image modifications.

2) Two dimensional image concatenation

This concatenation method combines the images together in a two dimensional layout. There are many ways to do this concatenation. We designed four methods of the kind of concatenation and named them by the way they concatenate the images as shown in figure 2.

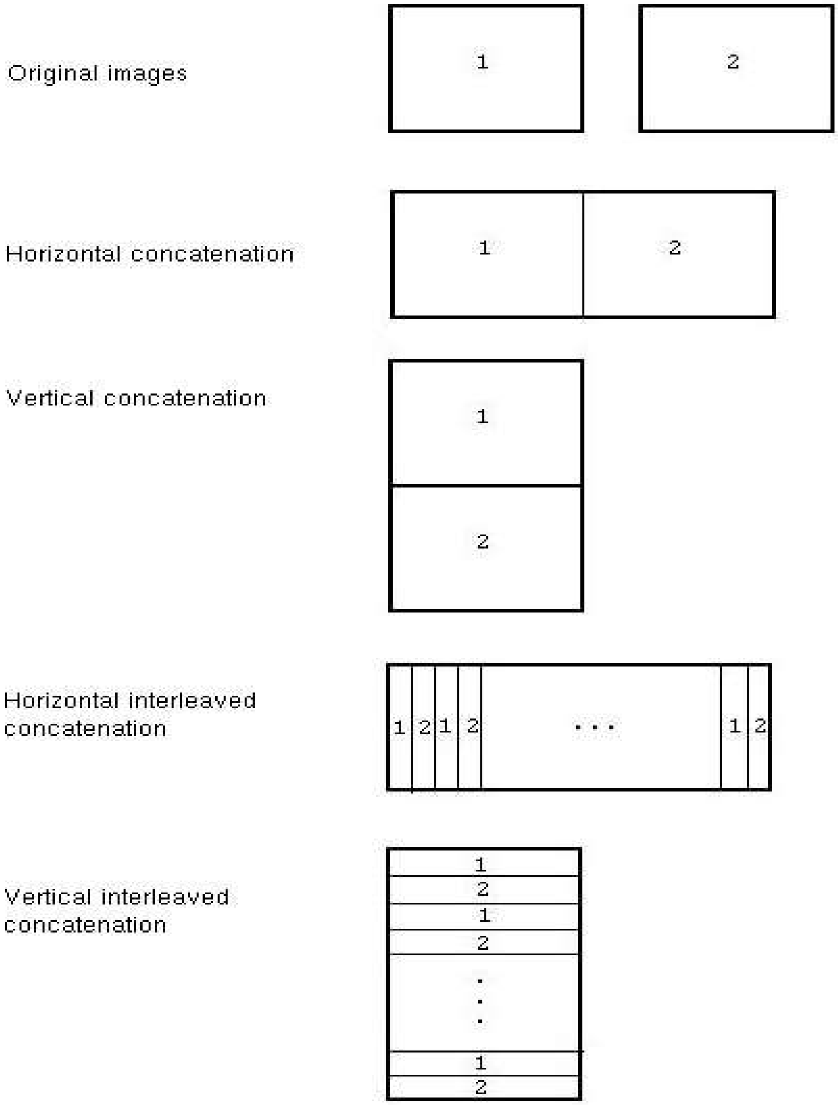

Figure 2. 2D image concatenation methods

Some image modifications are required, for example, in method 1, the horizontal image concatenation, the height of the images need to be the same, so we need to resize the images to have the same height. However, we do not want to change the image content as much as possible. As a result, we will maintain the image aspect ratio, so the image content in two dimension perspective would not change.

5.3 Implementing the experiments

Basically, we need to measure the normalized information distance between each images. There are two normalized information distance equations, so we will use both of them and compare the results.

Now we derive the equation (2.4) and (2.5) using equation (2.3) and we can see from the equation (2.3) that K(x,y) = K(xy) up to additive logarithmic precision [14], where xy is the concatenation of x and y. Therefore, we have two derived equation as

$$d_s(x,y)=1-\frac{K(x) + K(y) - K(xy)}{K(xy)} \tag{5.1}$$

$$d(x,y)=\frac{Max\{ K(xy)-K(y) , K(xy)-K(x) \}}{Max\{ K(x) , K(y) \}} \tag{5.2}$$

Now we use the compressors to measure each parameter as

K(x) is the size of the compressed image x, using the selected compressor.

K(y) is the size of the compressed image y, using the selected compressor.

K(xy) is the size of compressed image of x concatenated with y, using the selected concatenation method and compressor.

Afterwards, we do the statistic ranking based on the normalized information distance between each image and calculate the precision value. The precision values are defined as

$$\text{precision} = \frac{\text{number of correct returned images}}{\text{number of returned images}} \qquad (5.3)$$

5.4 Testing image data sets

In order to evaluate the proposed setup correctly, the appropriate image data set is required. We selected two image data sets for the experiment.

1) MIT texture image data set.

This image data set has 640 images that are classified into 15 classes. The images in this set that are in the same class are very similar to each other as shown in figure 3 and 4. All images in this set are in JPEG file format and have the same size.

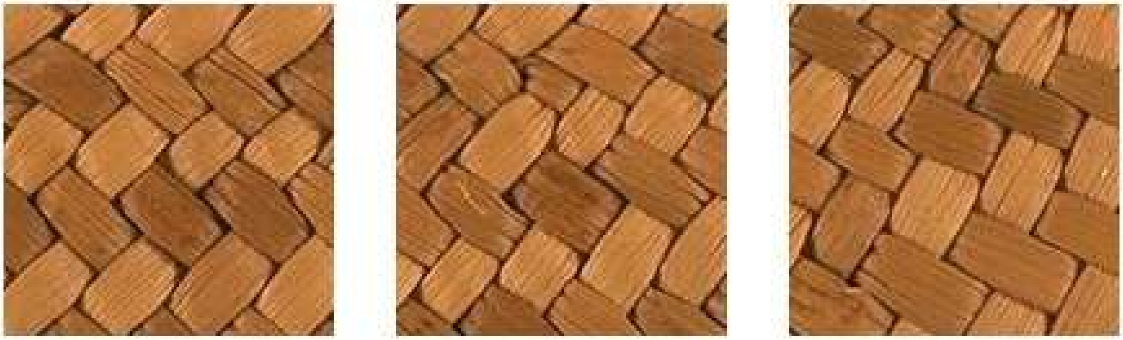

Figure 3. MIT texture image examples

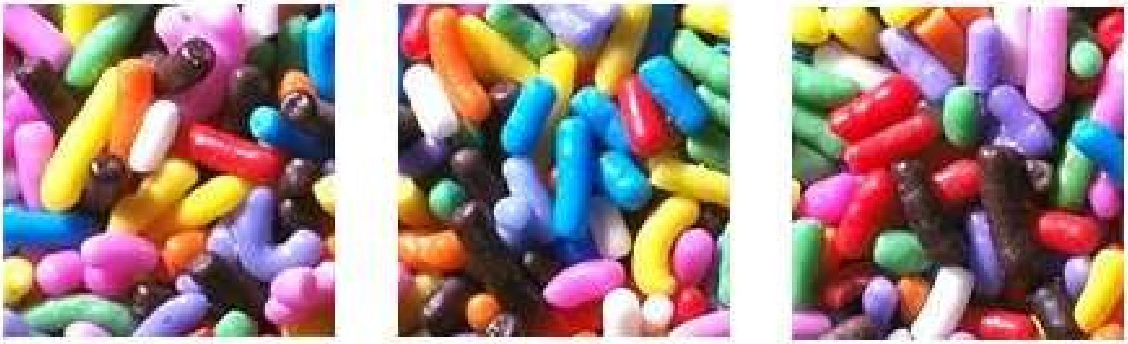

Figure 4. MIT texture image examples

2) Technical committee of IAPR image data set.

This set contains 1,000 images and comes with 30 standard queries. The standard query is the subset of the image data set that are classified to be in the same query. Therefore, an image can be in more than one standard query and also not all the images of the image data set are in the standard queries, unlike the MIT texture image data set. Also, the images that are in the same standard query are not as similar as in the MIT texture image data set, however, they are in the same category, such as animal, car, and sport. The example thumbnails of standard queries are shown in figure 5 and 6. All images in this set are in JPEG file format and have the a variety of sizes.

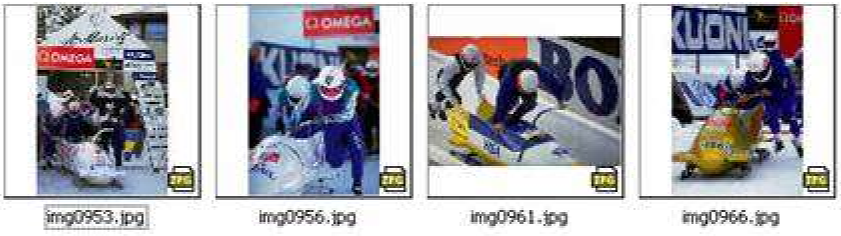

Figure 5. IAPR standard query example

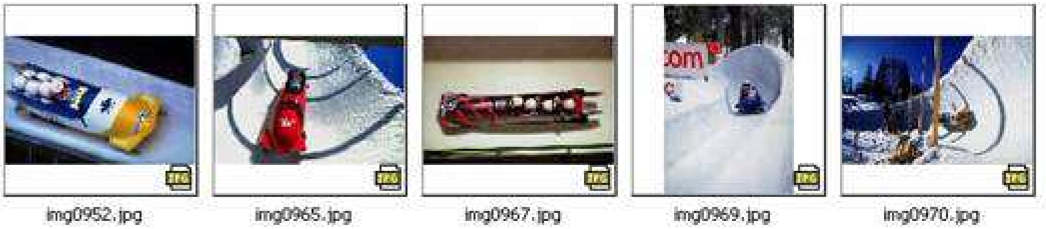

Figure 6. IAPR standard query example

5.5 McNemar's test

We use McNemar's test [22] to determine whether the image retrieval using approximated normalized information distance is statistically different from the random method.

	~B	B	
~A	n_{00}	n_{01}	X
A	n_{10}	n_{11}	Y
	I	J	Total

Table 1. McNemar's Test

In table 1, the symbols are defined as follow.

A = Approximated normalized information distance method classifier.

B = Random method classifier

n_{00} = The number of samples misclassified by A and B

n_{01} = The number of samples misclassified by A but not by B

n_{10} = The number of samples classified by A but not by B

n_{11} = The number of samples classified by A and B

I = The number of samples misclassified by B, equals to $n_{00} + n_{10}$

J = The number of samples classified by B, equals to $n_{01} + n_{11}$

X = The number of samples misclassified by A, equals to $n_{00} + n_{01}$

Y = The number of samples classified by A, equals to $n_{10} + n_{11}$

Total is the total number of samples, I + J = X + Y

In the experiments, the number of I, J, X,and Y are the results. Therefore, we can solve the equations for n_{00}, n_{01}, n_{10} ,and n_{11} . The z statistic is

$$z = \frac{|n_{01} - n_{10}| - 1}{\sqrt{n_{01} + n_{10}}} \tag{5.4}$$

The quantity z^2 is distributed approximately as chi square with one degree of freedom. Therefore, we can use z^2 value to test whether the classifiers have the same error rate.

6. Experimental results

The results can be separated into two sections as they are from the different image data sets. Also, there are some symbols that are used to represent the results. The symbols are defined as follow.

H-gzip	= Horizontal scan gzip compression
V-gzip	= Vertical scan gzip compression
JPEG2000	= JPEG2000 compression
H	= Horizontal concatenation
HI	= Horizontal interleaved concatenation
V	= Vertical concatenation
VI	= Vertical interleaved concatenation
$d_s(x,y)$	= Normalized information distance, equation 5.1
$d(x,y)$	= Normalized information distance, equation 5.2

The combination of compression and concatenation method is represent by compression method + concatenation method symbol, for example, H-gzip + V or HgV means horizontal scan gzip with vertical concatenation method.

In addition, for all tables in this section that have two values in one table cell, the above and below values represent the results from $d_s(x,y)$ and $d(x,y)$, respectively.

6.1 Experimental results on MIT image data set

The experiment on the MIT image data set is to measure the average image retrieval precisions. We compute the normalized information distances between one image to every other image in the data set. Therefore, we have the normalized information distances of every possible pair of the images. Afterwards, for each image, we retrieve 20 images that have the lowest normalized information distances to the image, and determine the image retrieval precisions at 20 images, also we compute the image retrieval precisions at 4, 8, 12, and 16 images.

Since all images in MIT data set have the same dimension, the sequences of the concatenated image contents that pass through the gzip compressor are the same in some cases. The combinations that yield the same results are listed as follow.

1) gzip + raw concatenation and H-gzip + V

2) H-gzip + H and H-gzip + VI

3) V-gzip + V and V-gzip + HI

Furthermore, we compare each combination method to the random method, between each combination method, and between the normalized information distances.

6.1.1 Image retrieval precisions

The results are shown in table 2, 3 and 4, then they can be plotted on the graphs for result comparisons as shown in figure 7 and 8.

	4	8	12	16	20
H	0.99 0.99	0.97 0.97	0.95 0.95	0.91 0.91	0.88 0.88
V	0.59 0.60	0.61 0.61	0.62 0.62	0.61 0.61	0.60 0.60
HI	0.93 0.92	0.89 0.88	0.86 0.85	0.82 0.80	0.77 0.73

Table 2. Image retrieval precisions at different numbers of images using horizontal scan gzip with the concatenation methods

	4	8	12	16	20
H	0.51 0.52	0.55 0.55	0.57 0.58	0.59 0.59	0.58 0.59
V	0.99 0.99	0.97 0.97	0.94 0.94	0.91 0.91	0.88 0.88
VI	0.94 0.92	0.90 0.89	0.87 0.85	0.83 0.81	0.78 0.77

Table 3. Image retrieval precisions at different numbers of images using vertical scan gzip with the concatenation methods

	4	8	12	16	20
H	0.98 0.98	0.97 0.97	0.95 0.95	0.92 0.92	0.89 0.89
V	0.97 0.97	0.94 0.95	0.92 0.93	0.89 0.89	0.86 0.86
HI	0.65 0.54	0.60 0.49	0.57 0.47	0.54 0.45	0.51 0.44
VI	0.54 0.45	0.51 0.43	0.50 0.42	0.48 0.41	0.47 0.40

Table 4. Image retrieval precisions at different numbers of images using JPEG2000 with the concatenation methods

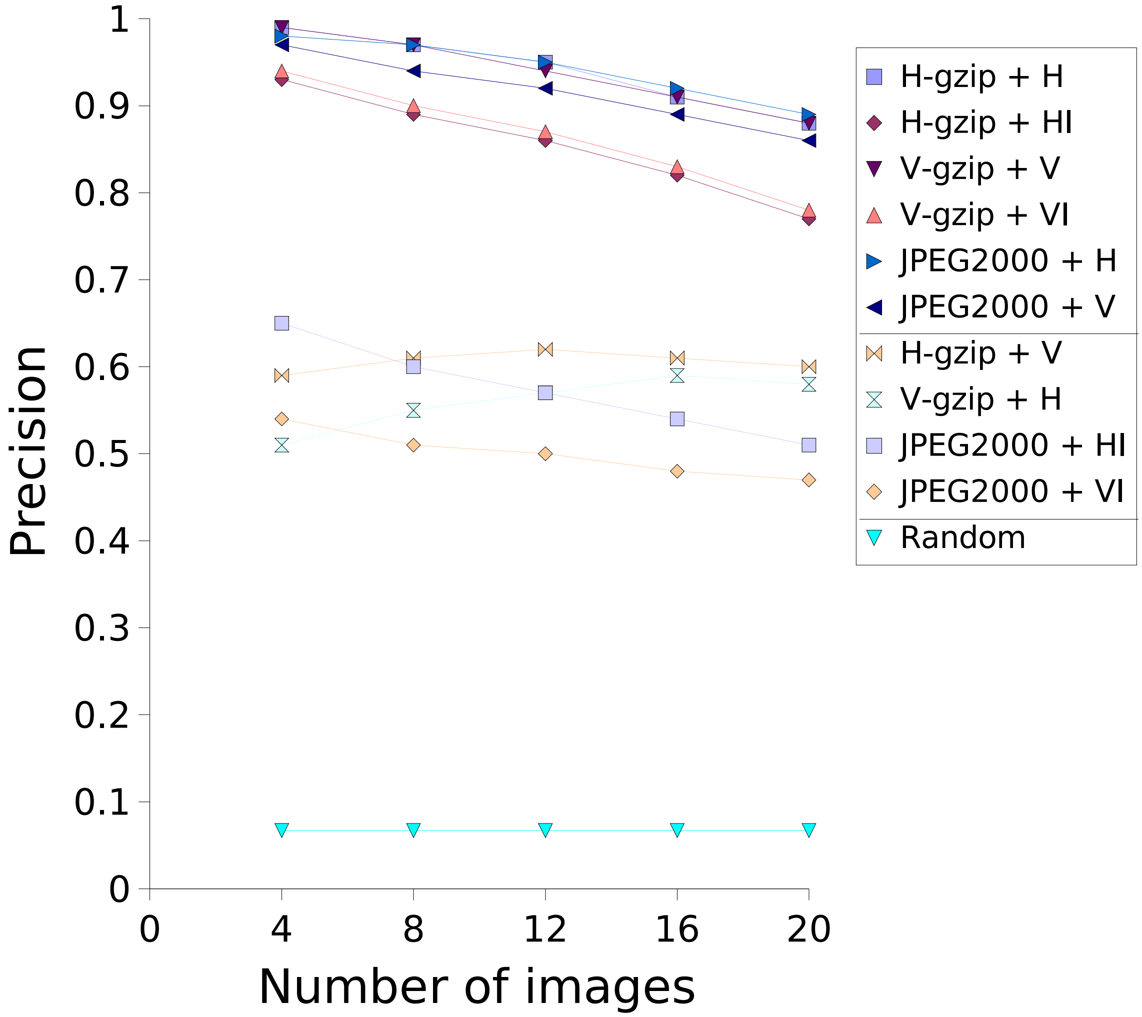

Figure 7. Image retrieval precisions at different numbers of images using $d_s(x,y)$ on MIT image data set

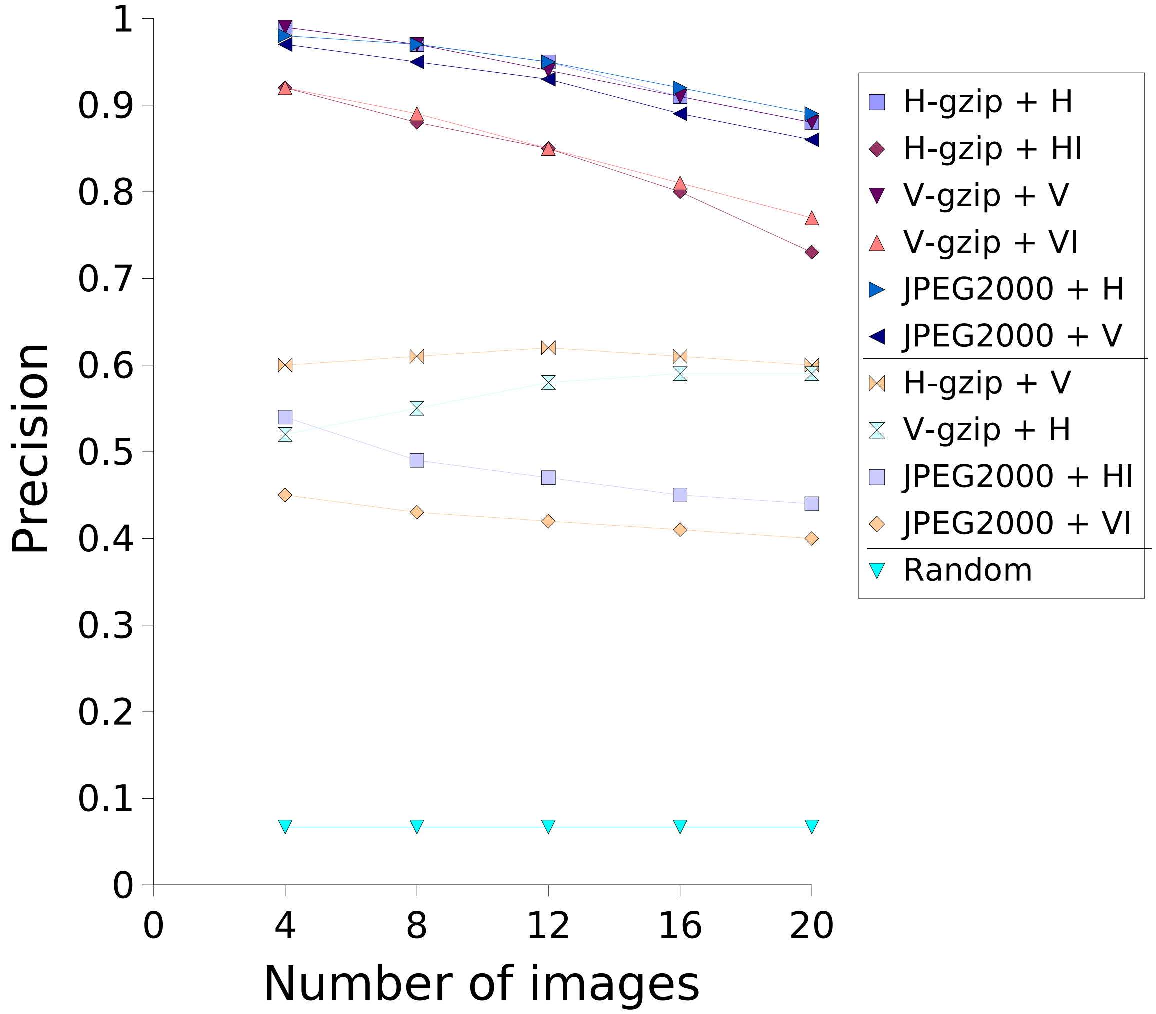

Figure 8. Image retrieval precisions at different numbers of images using d(x,y) on MIT image data set

The results from both normalized information distances have the same pattern. The image retrieval precisions from the graphs can be separated into two groups.

1) High precision range (0.7 to 1)

- H-gzip + H (H-gzip + VI)
- H-gzip + HI
- V-gzip + V (V-gzip + HI)
- V-gzip + VI
- JPEG2000 + H
- JPEG2000 + V

2) Low precision range (0.4 to 0.7)

- H-gzip + V (gzip + raw concatenation)
- V-gzip + H
- JPEG2000 + HI
- JPEG2000 + VI

In gzip compression case, the results show that the combinations that are in the high precision range have the mixture of the two images in the compression sequences, the order of the pixels of the concatenated image that is processed through the compressor. The combinations of concatenation and compression methods in the low precision range have no mixture of images in the compressing sequences, that is, the compressor scan through the first image until the end of the image, then process the second image. Since the gzip compressor has small window size, the mixture of images in the compression sequence effects the precision of the image retrievals considerably.

In JPEG2000 compression case, the results show that using JPEG2000 compression without interleaved concatenations yields higher precision than with the interleaved concatenations. Although the interleaved concatenation method moves the contents of the two images that are in the same area closer, JPEG2000 works better with non-interleaved concatenation that perspectively maintain the original images. A block based interleaving may work better for JPEG2000, since it would maintain and mix wavelet block better, future research could test this conjecture.

The compressors that we use have different behaviors and techniques. Therefore, using the right concatenation methods with the right compression methods yield the better results.

6.1.2 Comparisons between each combination method and the random method

We compare each combination method to the random method by computing the z values using McNemar's test. The random method has the probability of retrieving the corrected classified images of 1/15 (15 classes of images).

	H-gzip	V-gzip	JPEG2000
H	100.97 100.86	77.33 77.77	102.33 102.00
V	78.58 78.79	100.82 100.71	99.43 99.74
HI	92.93 91.56	Same as V-gzip + V	70.75 62.78
VI	Same as H-gzip + H	93.88 92.55	65.87 59.04

Table 5. z values from McNemar's test between each combination method and the random method

A table of critical values for chi square shows that with one degree of freedom the critical value of chi square is 3.84 at 0.05 level of significance. That is, the critical value of z from the McNemar's test is 1.96 at 0.05 level of significance. In table 5, the z values show that all combinations are statistically different from the random method and perform better than the random method for this image data set.

6.1.3 Comparisons between each combination method

We also use the same comparison method in section 6.1.2, computing the approximated chi square value using McNemar's test to compare between each combination method and the result z values are shown in table 6. The values are the results of comparing the two methods respective to the row and column.

		H-gzip			V-gzip			JPEG2000			
		H	V	HI	H	V	VI	H	V	HI	VI
H-gzip	H										
	V	56.29 55.59									
	HI	36.30 39.04	33.74 29.33								
V-gzip	H	65.10 64.41	22.04 21.94	45.35 41.20							
	V	1.57 1.57	56.00 55.23	35.38 38.16	64.77 64.05						
	VI	33.36 36.40	36.23 31.91	7.39 7.19	46.99 42.88	32.85 35.89					
JPEG 2000	H	4.72 4.14	59.07 58.63	30.74 34.29	66.78 66.51	5.34 5.75	28.46 32.19				
	V	6.43 4.80	55.12 55.45	21.19 26.19	62.45 62.77	5.74 4.09	18.66 23.82	14.36 13.17			
	HI	65.37 72.32	16.96 30.64	48.61 57.32	2.32 17.51	64.82 71.86	50.00 58.24	66.81 74.84	62.18 71.09		
	VI	66.01 71.43	25.81 35.92	52.72 57.47	11.42 22.68	65.44 70.77	54.66 59.01	72.33 77.91	67.93 74.31	11.80 8.44	

Table 6. z values from McNemar's test between each combination method

The results show that most of the methods are statistically different at 0.05 level of significance except for V-gzip + V and H-gzip + H on both normalized information distances

6.1.4 Comparisons between the normalized information distances

In order to answer the hypothesis question, we compare the results from the same combination of concatenation and compression methods, but using the different normalized information distances. We use the same comparison method in section 6.1.2, computing the approximated chi square value using McNemar's test to compare between each result and the result z values are shown in table 7.

H-gzip	H	2.05
	V	3.91
	HI	14.29
V-gzip	H	3.51
	V	2.2
	VI	14.35
JPEG2000	H	0.25
	V	2.82
	HI	29.82
	VI	26.51

Table 7. z values from McNemar's test between each normalized information distances

From table 7, we can display the results in the figure 9.

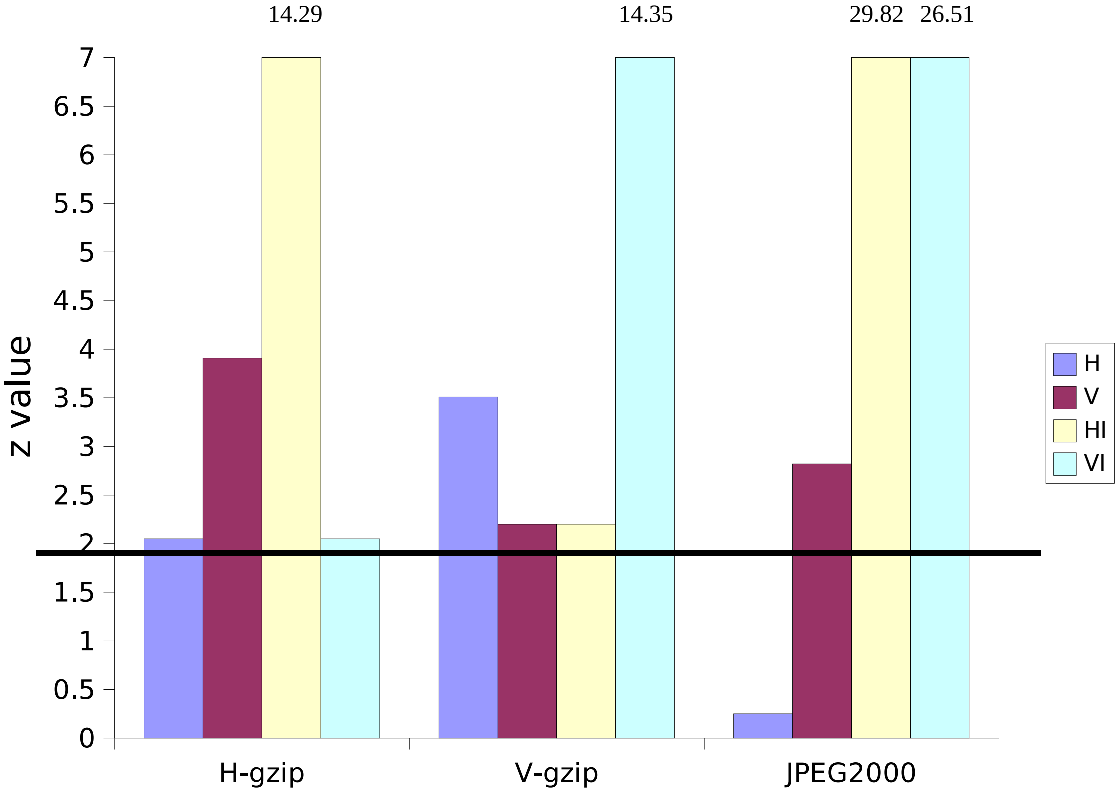

Figure 9. z values between each normalized information distances

Since the comparison methods are the same, the critical value of z from the McNemar's test is 1.96 at 0.05 level of significance. From table 7, the normalized information distances are statistically different in most of the combination of concatenation and compression methods except in JPEG2000 + H. However, note that the z values of the combination of concatenation and compression methods that are in the high precision range in section 6.1.1 are very close to the critical value of the z value, unlike the combination of concatenation and compression methods that are in the low precision range.

6.2 Experimental results on IAPR image data set

The experiment on the IAPR image data set is different from the experiment on the MIT image data set, because IAPR image data set classifies images differently. IAPR image data set has 30 standard queries, which have different numbers of images in the queries, range from 1 to 12 images. We designed two experiment setups on the IAPR image data set.

In the first experiment, we use the standard queries that have only 2 images in the queries. There are 7 standard queries that are in the specification. For each image, perform the image retrieval and record the rank of the corrected classified image, its pair in the query. Afterwards, we average the recorded ranks for each combination of concatenation and compression method. Also, the results are compared to the random method whether they are statistically different.

The second experiment performs on 3 selected standard queries. The selected standard queries must have more than 2 images in the queries. For each standard query, each image in the query is used to retrieve the other images in the query using raw concatenation and gzip compression method and record the average ranks of the first lowest ranks, the second lowest ranks and so on. Also, the results are compared to the random method whether they are statistically different.

6.2.1 Results on 2 images standard queries

Unlike MIT image data set, the images in IAPR image data set are different in dimensions. Therefore the results from each combination of concatenation and compression method would be different. The 2 images standard queries example thumbnails are shown in figure 10, 11, and 12. The results are shown in table 8.

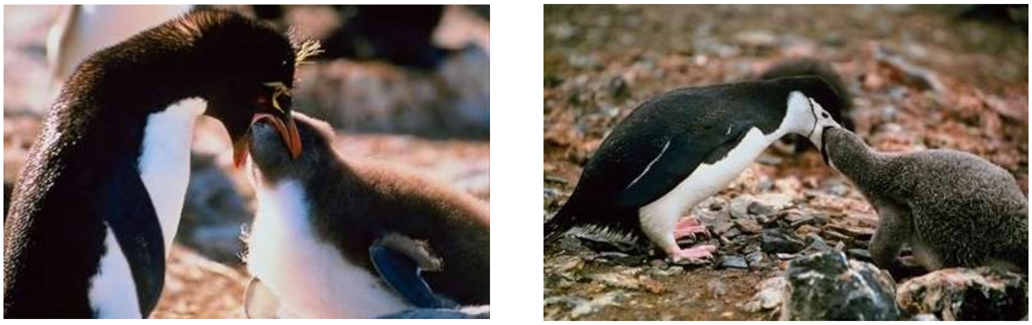

Figure 10. IAPR standard query number 5

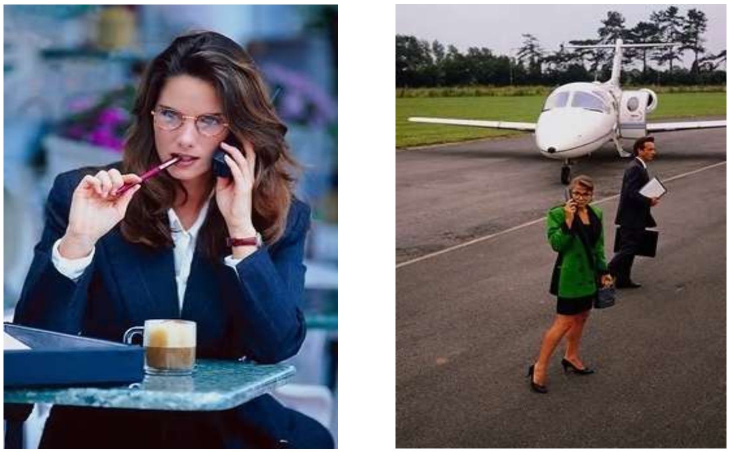

Figure 11. IAPR standard query number 25

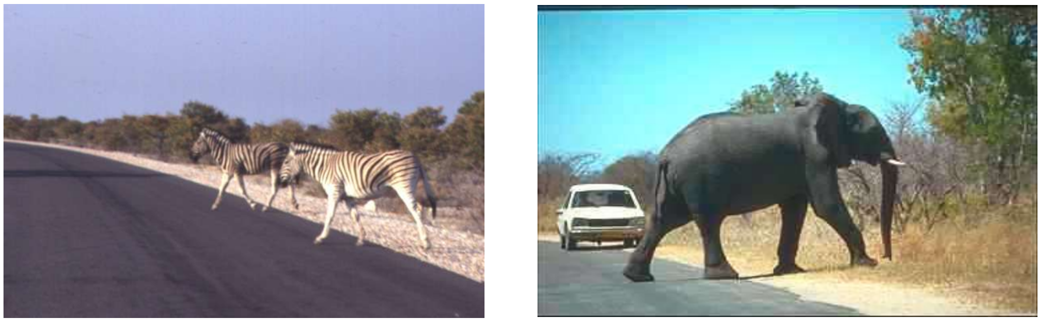

Figure 12. IAPR standard query number 28

	H-gzip	V-gzip	JPEG2000
Raw	486.29 487.86		
H	771.36 750.86	665.50 665.00	735.21 733.43
V	510.64 526.57	719.57 696.71	780.43 757.14
HI	707.50 718.21	756.71 749.86	503.71 519.64
VI	692.86 697.43	633.21 626.36	665.21 506.00

Table 8. Average ranks of image retrievals for each combination method on 2 images standard queries.

In order to determine whether each combination of concatenation and compression method and random method are statistically different, we use a significance test which is known as z-test, where the test statistic is defined as $z=\frac{\bar{x}-u_0}{sd/\sqrt{n}}$, where

$\bar{x}$ = mean of samples

u_0 = mean of population

sd = standard deviation of population

n = number of samples

In this case the rank is range from 1 to 999. Therefore, the random method is the population in this case and normally distributed with the average rank of 500 and 288.39 standard deviation. After calculating the z values, we measure the P-value, possibility of observing extreme values which imply that they are likely to be different. Since the null hypothesis is the mean of the results equal to the random method results, we use two-sided test to compute the P-value to test against the null hypothesis as shown below.

$P(H_a : \bar{x} \neq u_0) = 2P(Z \geq |z|)$ in standard normal distribution (6.1)

The result P-value are shown in table 9.

	H-gzip	V-gzip	JPEG2000
Raw	0.8572 0.8650		
H	0.0000 0.0012	0.0316 0.0324	0.0022 0.0024
V	0.8886 0.7338	0.0046 0.0108	0.0000 0.0008
HI	0.0072 0.0046	0.0008 0.0012	0.9620 0.8026
VI	0.0124 0.0104	0.0854 0.1010	0.0324 0.9362

Table 9. P-values of each combination and random method comparison on 2 images standard queries

In figure 13, the average ranks results are shown with the minimum ranks, shown as green dots, and maximum ranks, shown as red dots. Also, the critical P-values respective to the rank are shown with red lines.

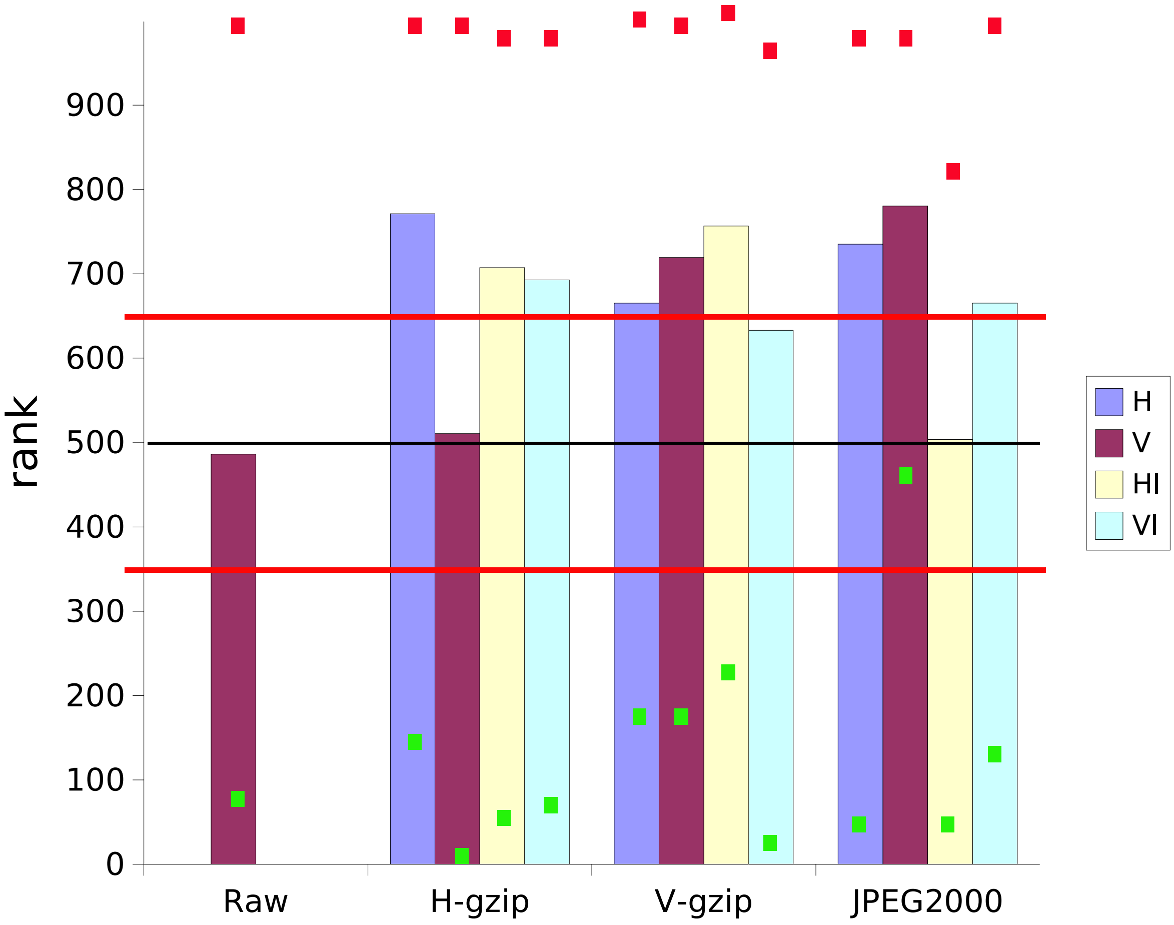

Figure 13. Average ranks of image retrievals for each combination method on 2 images standard queries

At the significant level of 0.05, most of the combinations of concatenation and compression method are statistically significant from the random method and perform worse than the random method except the following combinations.

1) gzip + raw on both normalized information distances

2) H-gzip + V on both normalized information distances

3) V-gzip + VI on both normalized information distances

4) JPEG2000 + HI on both normalized information distances

5) JPEG2000 + VI on d(x,y)

6.2.2 Results on the selected standard queries

The average rank results of image retrievals using raw concatenation with gzip compression on standard query number 3, 7, and 14 are shown in table 10, 11, and 12, respectively. The example thumbnails of standard query number 3, 7, and 14 are shown in figure 14, 15, and 16, respectively. Afterwards, the results from all queries and random method result are shown in figure 17. The result analysis in this section is left for the future work.

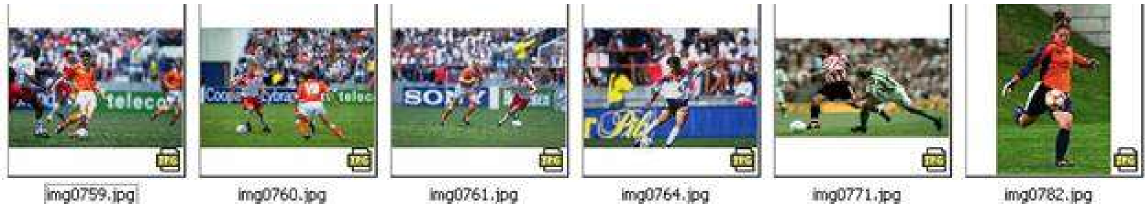

Figure 14. IAPR standard query number 3

Image return order	Average rank
1	221.83 207.00
2	353.33 353.50
3	419.50 428.00
4	571.83 572.83
5	626.83 660.83

Table 10. The average rank result on IAPR standard query number 3 using raw concatenation method with gzip compression

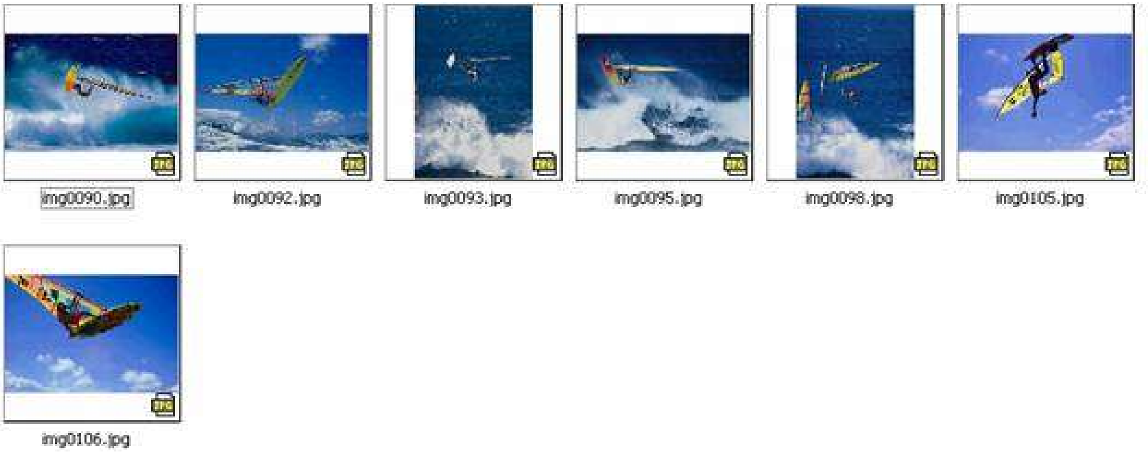

Figure 15. IAPR standard query number 7

Image return order	Average rank
1	207.14 183.00
2	352.71 337.29
3	432.71 400.29
4	496.86 463.29
5	581.86 557.29
6	797.14 781.29

Table 11. The average rank result on IAPR standard query number 7 using raw concatenation method with gzip compression

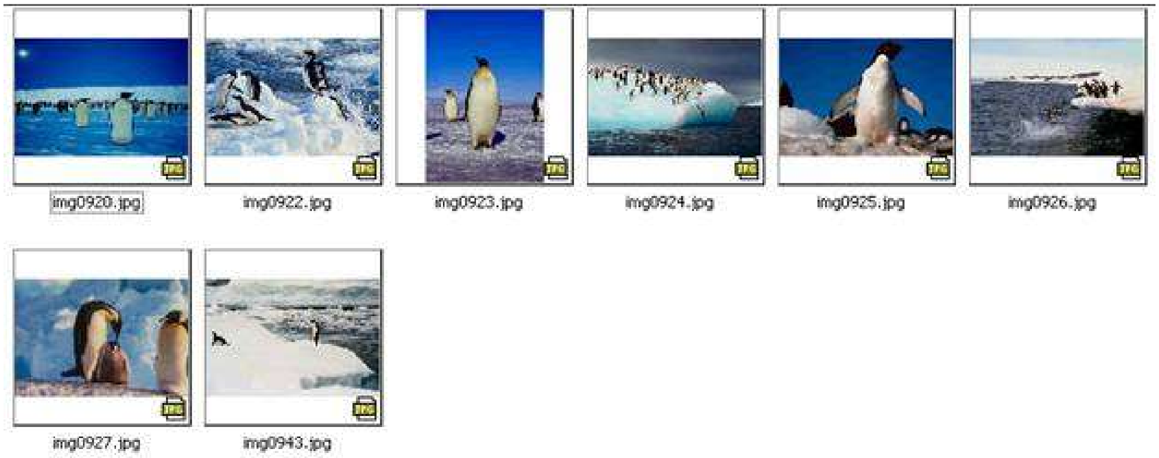

Figure 16. IAPR standard query number 14

Image return order	Average rank
1	115.25 98.63
2	178.88 170.25
3	251.00 230.75
4	494.38 463.00
5	620.38 599.75
6	741.63 731.00
7	838.75 827.75

Table 12. The average rank result on IAPR standard query number 14 using raw concatenation method with gzip compression

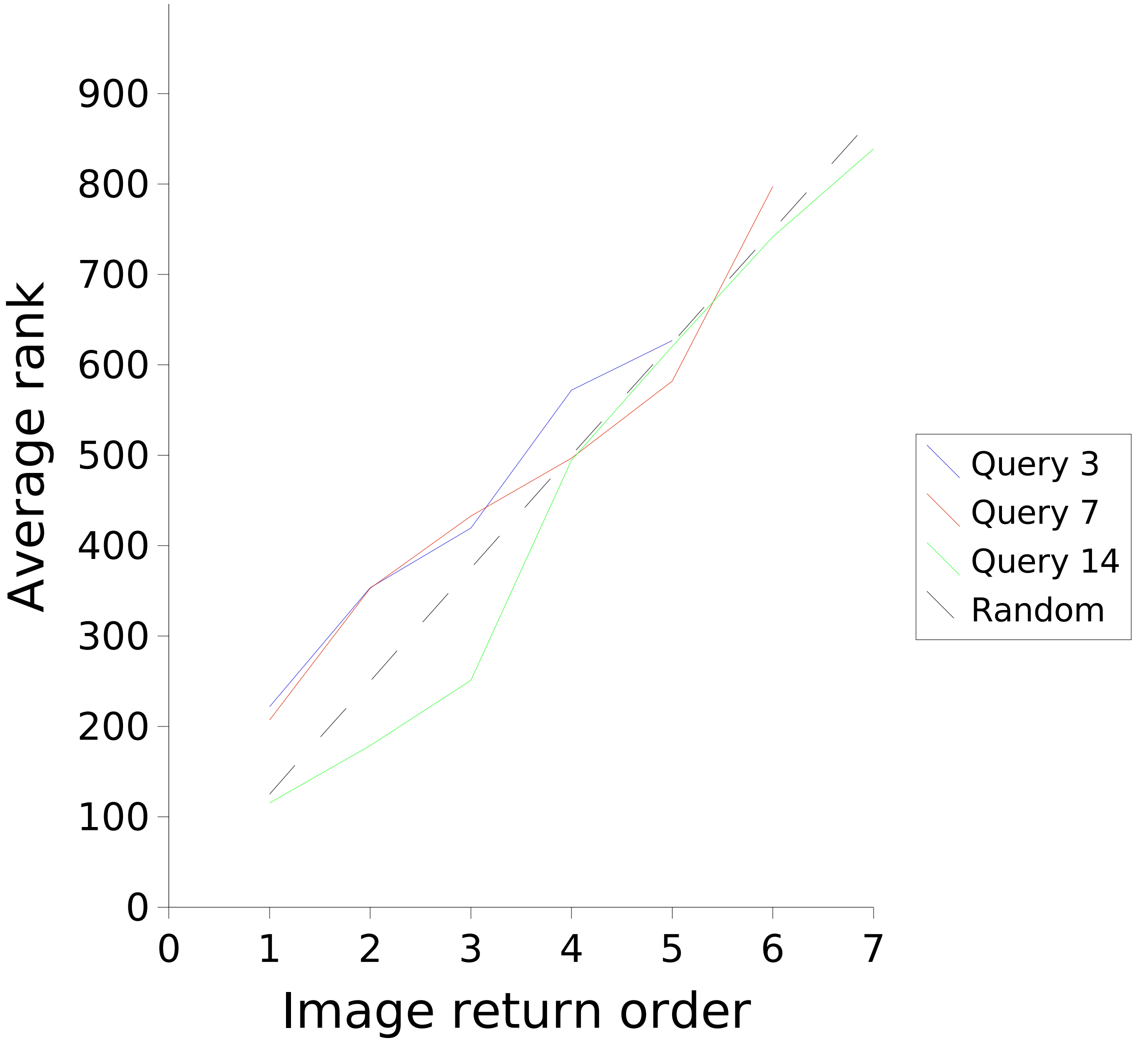

Figure 17. Standard queries and random method comparison

7. Conclusion

A mathematical theory of similarity distances has been developed and shown that there is a universal similarity distance, the normalized information distance by Li et al. [14]. The normalized information distance is based on the noncomputable notion of Kolmogorov complexity. Even so, the experiments on the theory with Kolmogorov complexity approximation methods have shown the remarkable success. As a result, we attempt to use the normalized information distance as a measurement in image retrieval problem. In this paper, we approximate the Kolmogorov complexity by using compression methods and image concatenation methods. The proposed approach was able to generate statistically significant results that are better than the random method on MIT texture image data set, which its images that are in the same class are very similar. In most cases, the different combinations of concatenation and compression methods yield statistically different results, and different normalized information distances also yield statistically different results. On the IAPR image data set, the proposed approach was not able to generate a good result compare to the random method due to the image classification in the image data set that is determined by what that image is about, which is harder to capture.

The possible future work would be exploring other methods of approximating the Kolmogorov complexity, concatenation and compression methods, that may yield the better result. Another possible future work would be using the normalized information distance on the regions and feature vectors of the images.

Bibliography

[1] C. Boncelet. Simple, High Performance Lossless Image Compression. *Image Processing, 2001. Proceedings. 2001 International Conference,* 7-10 Oct. 2001, 498-501 vol.3.

[2] N. Chater and P. M. B. Vitanyi. The Generalized Universal Law of Generalization. University of Warwick CWI and Universiteit van Amsterdam. *Journal of Mathematical Psychology,* 47:3(2003), 346-369.

[3] C. Chen. On the Selection of Image Compression Algorithms. *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference* on, 16-20 Aug. 1998, 1500-1504 vol.2.

[4] X. Chen, M. Li, Brian McKinnon, and Amit Seker. A Theory of Uncheatable Program Plagiarism Detection and Its Practical Implementation. SID Website at http://dna.cs.ucsb.edu/SID/, Manuscript available at http://www.cs.ucsb.edu/~mli/sid.ps, 2002.

[5] Z. Chi and J. Kong. Image Content Classification Using a Block Kolmogorov Complexity Measure. *Signal Processing Proceedings, 1998. ICSP '98. 1998 Fourth International Conference*, 12-16 Oct. 1998, 1185-1188 vol.2.

[6] R. Fleix, R. Albersmann, and S. Reddig. Industrial Application of Fuzzy I mage Comparison in Quality Control. *Proceedings of the Sixth IEEE International Conference*, 1-5 July 1997, 1405-1409 vol.3.

[7] M. A. Golner, W. B. Mikhael, V. Krishnan, and A. Ramaswamy. Region Based Variable Quantization for JPEG Image Compression. *Circuits and Systems, 2000. Proceedings of the 43rd IEEE Midwest Symposium* on, 8-11 Aug. 2000, 604-607 vol.2.

[8] gzip compression. URL http://www.gzip.org.

[9] D. P. Huttenlocher, W. J. Rucklidge and G.A. Klanderman. Comparing Images Using the Hausdorff Distance Under Translation. *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference*, 15-18 June 1992, 654-656.

[10] D. J. Jackson and S. J. Hannah. Comparative Analysis of Image Compression Techniques. *System Theory, 1993. Proceedings SSST '93., Twenty-Fifth Southeastern Symposium,* 7-9 March 1993, 513 -517.

[11] J. Jiang. Neural Network Technology for Image Compression. *Broadcasting Convention, 1995. IBC 95., International* , 14-18 Sep 1995, 250-257.

[12] R. Jin and A. G. Hauptmann. Using a Probabilistic Source Model for Comparing Images. *Image Processing. 2002. Proceedings. 2002 International Conference*, 24-28 June 2002, 941-944 vol.3.

[13] JJ2000, Java Implementing of JPEG 2000. URL http://jpeg2000.epfl.ch.

[14] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi. The Similarity Metric. *Proc. 14th ACM-SIAM Symp*. Discrete Algorithms, 2002.

[15] M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications,* Springer-Verlag, New York, 2nd Edition, 1997.

[16] T. Markas and J. Reif. Multispectral Image Compression Algorithm. *Data Compression Conference, 1993. DCC '93.*, 30 March-2 April 1993, 391-400.

[17] A. J. Mikhail. *Algorithms and Theory of Computation handbook.* CRC Press LLC, Florida, 1999, 29-18 to 29-20.

[18] N. Ranganathan, S. G. Romaniuk, and K. R. Namuduri. A Lossless Image Compression Algorithm Using Variable Block Size Segmentation. *Pattern Recognition, 1994. Vol. 3 - Conference C: Signal Processing, Proceedings of the 12th IAPR International Conference*, October 9-13, 1994, 40-44 vol.3.

[19] N. Sahasrabudhe, J. E. West, R. Machiraju, and M. Janus. Structured Spatial Domain Image and Data Comparison Metrics. *Visualization '99. Proceedings*, 24-29 Oct. 1999, 97-515.

[20] M. Schmidr, V. Kreinovich, and L. Longpre. Kolmogorov Complexity-Based Ideas for Locating Text in Web Images. *Circuits and Systems, 1999. 42nd Midwest Symposium*, 8-11 Aug. 1999, 543-546 vol. 1.

[21] Q. Wang, Z. Chi, and R. Zhaof. Hierarchical Content Classification and Script Determination for Automatic Document Image Processing. *Pattern Recognition, 2002. Proceedings. 16th International Conference*, 11-15 Aug. 2002, 77-80 vol.3.

[22] A. Webb. *Statistical Pattern Recognition*. John Wiley and Sons, LTD., Hoboken, NJ, 2nd edition, 2002.

[23] S. K. Yip and Z. Chi. Page Segmentation and Content Classification for Automatic Document Image Processing. *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium*, 2-4 May 2001, 279-282.

VITA

Supat Supamahitorn

Candidate for the Degree of

Master of Science

Thesis: INVESTIGATION OF A KOLMOGOROV COMPLEXITY BASED SIMILARITY METRIC FOR CONTENT BASED IMAGE RETRIEVAL

Major Field: Computer Science

Biographical:

Personal Data: Born in Bangkok, Thailand, On November 27, 1979, the son of Udom and Penpun Supamahitorn

Education: Graduated from Suankularb College, Bangkok, Thailand, in March 1997; received Bachelor of Engineering degree in Computer Engineering from Chulalongkorn University, Bangkok, Thailand, in April 2001. Completed the requirements for the Master of Science degree with a major in Computer Science at Oklahoma State University in December 2004.

Name: Supat Supamahitorn Date of Degree: December, 2004

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: INVESTIGATION OF A KOLMOGOROV COMPLEXITY BASED SIMILARITY METRIC FOR CONTENT BASED IMAGE RETRIEVAL

Pages in Study : 39 Candidate for the Degree of Master of Science

Major Field: Computer Science

This paper introduces an image retrieval approach using normalized information distance based similarity metric to determine the difference between the images. The similarity metric is based on Kolmogorov complexity and measures the amount of shared information between images. Although the Kolmogorov complexity is uncomputable, we are following Vitanyi's approach for approximating it.

Advisor's Approval: Dr. Douglas R. Heisterkamp