An Experimental Comparison of Feature Extraction and Distance Metrics for Image Retrieval

Ramon F. Pessoa, William R. Schwartz, Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Minas Gerais, Brazil, 31270-901
Email: {ramon.pessoa, william, jefersson}@dcc.ufmg.br

***Abstract*—This paper seeks to do a comparative study of different features and distance metrics in order to analyze the impact of these factors in the process of Content-Based Image Retrieval (CBIR). One of the main contributions of this work was statistically analyze the impact of distance metrics in the process of image retrieval by content. We also observed statistically the impact of the variability among different classes and also the variability between images of the same image class. The results showed, for a sample collected, that the variation attributed to the class is approximately 99.85%. This confirms the fact that each algorithm will work best in a given situation. The comparative study showed the algorithms which had better accuracy rate to recover different image classes (in the dataset analysed) and also presented the reasons that possibly made these algorithms had better accuracy rate.**

***Keywords*-CBIR experimental comparison; statistical analysis; feature extraction algorithms; distance metrics;**

I. Introduction

The term CBIR (Content-Based Image Retrieval) describes the process of retrieving desired images from the large collection of database on the basis of features that can be automatically extracted from the images. The ultimate goal of a CBIR system is to avoid the use of textual descriptions in the hunt for an image by the user.

In CBIR, retrieval of image is based on similarities in their contents, i.e., textures, colors, shapes etc., which are considered the lower level features of an image. In CBIR each image stored in the database, has its features extracted and compared to the features of the query image. Thus, broadly, it involves two processes, feature extraction and feature matching [1]. The search is usually based on similarity rather than on exact match and the retrieval results are then ranked accordingly to a similarity index.

This paper aims to assess the impact of different image features and distance metrics in CBIR process. Several algorithms for image recovery content were implemented using different distance metrics in order to show whether the accuracy of relevant image retrieval really is dramatically impacted by these factors and what are noticeably impacting factors.

II. Related Work

Image retrieval and content-based image retrieval (CBIR) are well-known fields of research in information management in which a large number of methods have been proposed and investigated but in which still no satisfying general solutions exist [2].

Bao et al., [3], investigated eight similarity measures through some remote sensing image retrieval. From the experiment results it can be found that X^2 statistical distance measure and cosine of the angle measure perform better than others.

Beecks et al., [4], study the behavior of the similarity measures by discussing their properties. They compared the Hausdorff Distance, Perceptually Modified Hausdorff Distance, Weighted Correlation Distance, Earth Movers Distance, Signature Quadratic Form Distance and evaluated experimental results with respect to their qualities of effectiveness, as well as efficiency.

Rubinstein et al., [5], present a comprehensive perceptual study and analysis of image retargeting. They present analysis of the users responses, where they find that humans in general agree on the evaluation of the results and show that some retargeting methods are consistently more favorable than others.

None of those studies evaluated Feature Extraction algorithms and Similarity Measures using the statistical methods described in this paper.

III. Methodology

The CBIR system implemented in this work uses four feature extraction algorithms. Three color algorithms (HSV, RGB, YUV) and a new algorithm, proposed in this work, which consists in reducing a certain image for a range of 8 (Sampling-8)[1]. The similarity measures used were: 1) Cosine, 2) Euclidean, 3) Manhattan and 4) Chessboard.

To achieve the objectives of this work, the following activities were performed:

1) We performed a comparative analysis between four simple feature extraction algorithms with images of 10 different classes of images and a comparative analysis of the best simple algorithm with a combined approach of these four algorithms.
2) An analysis of the impact of four distance metrics to evaluate the impact of them in CBIR, where we used a statistical method called one-factor design.

[1] The downscaling by a factor of 8 was empirically defined.

3) To confirm the existing variability between images of the same class we performed a 2^2 factorial design to evaluate the influence of the distance metric and class of images on the accuracy of the image retrieval process.
4) We performed a new comparative analysis between the same four simple feature extraction algorithms (Sampling-8, HSV, RGB, YUV) but now only in images of the same class of images, this allowed a characterization of each of these algorithms.

A. *Database used*

We have used a standard database for testing, the WANG database [6]. WANG database is a subset of 1,000 images which form 10 classes of 100 images each. The images classes are: 1) Africa; 2) Beach; 3) Monuments; 4) Buses; 5) Dinosaurs; 6) Elephants; 7) Flowers; 8) Horses; 9) Mountains; 10) Food. Given a query image, it is assumed that the user is searching for images from the same class, and the remaining 99 images from the same class are considered relevant and the images from all other classes are considered irrelevant.

B. *Environment*

The tests were performed using a laptop Dell XPS 15 L502X, Intel Core i7-2670QM processor, 2.2GHz Turbo Boost 2.0 of 3.1 Ghz, 6Mb Cache, 8 Threads, Windows 7 Professional 64-bit, 6 GB of memory, DDR3, 1333MHz (1x2Gb+1x4Gb).

IV. Results and Discussion

In this section, we describe a comparative analysis of four feature extraction algorithms (Sampling-8, HSV, RGB, YUV) and four similarity measures (Cosine, Euclidean, Manhattan, Chessboard). We also analyse a combined approach of feature extraction (Sampling-8 + HSV + RGB + YUV). The comparative analysis is the evaluation of the runtime and accuracy of these algorithms.

A. *Determining the Sample Size*

To carry out paired observations, the first step is to determine the sample size for the results to have statistical validity. The statistical formula for determining the sample size considering the amount of preliminary observations (n), the arithmetic mean (x), standard deviation (s) results, the desired confidence (p), the error rate considered (r) and the value of the student-t distribution related (for sample less than 30 observation) [7]:

$$Size_{(Sample)} = \sqrt{\frac{100 * t_{[p;n-1]} * s}{r * \bar{x}}} \quad (1)$$

Initially, all simple feature extraction algorithms (Sampling-8, HSV, RGB, YUV) and the combined one was set to use the cosine distance metric.

For the metric "Time", we run 3 times for each image of a specific class and applying the abovementioned formula we found we had to run 9.85 times (10 times).

For the metric "Precision", we run 10 times for each image of a specific class and applying the abovementioned formula we found we had to run 34.46 times (35 times). Therefore, based on conservative statistical approach was adopted 35 observations as the sample size for paired observation.

B. *Confidence Interval and Bonferroni Correction*

After 35 replications, we calculate the confidence interval for the "Time" (See Figure 1) and "Precision" (See Figure 2) metrics with 99% of confidence to the four simple algorithms: HSV, RGB, YUV, 8-Sampling.

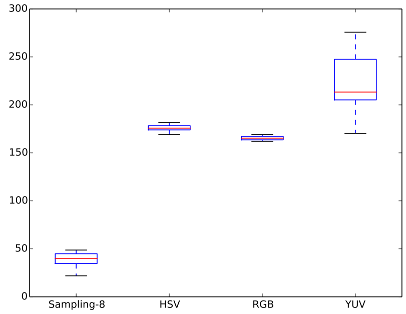

Fig. 1. Confidence interval for the metric "Time" with 99% confidence for the four simple algorithms: HSV, RGB, YUV, 8-Sampling

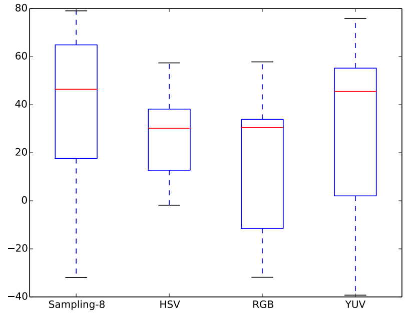

Fig. 2. Confidence interval for the metric "Precision" with 99% confidence for the four simple algorithms: HSV, RGB, YUV, 8-Sampling

Our goal in this analysis is to make paired comparisons of the "best simple algorithm" with "the combined approach algorithm of the four algorithms".

As we can see in Figure 1, the sampling-8 algorithm is noticeably better than the other algorithms in relation to the "Time". But in relation to "Precision" (Figure 1), the confidence interval for each of the algorithms has the mean of another algorithm, so we can not conclude that they are different.

To confirm that these algorithms are not really different from each other we use the Bonferroni Correction that is a method used to counter act the problem of multiple comparisons. Bonferroni Correction tests each pair with a significance level of $\binom{\alpha}{k}$, where k is the number of comparisons and α is the significance level. We choose a significance level of 10% (Confidence Interval = 90%).

Bonferroni correction shows that Sampling-8, HSV, RGB, YUV are not significantly different with defined confidence. But Sampling-8 algorithm is faster, so, we choose the

Sampling-8 approach to compare with the Combined Approach.

C. Comparing Two Alternatives

In this subsection we show the results of paired observations of the following algorithms:

- Algorithm combined (A): Sampling-8 + HSV + RGB + YUV using cosine distance.
- Best single algorithm (B): Sampling-8 using cosine distance.

The analysis of paired observation is straightforward. The two samples are treated as one sample of n pairs. For each pair, the difference in performance can be computed. A confidence interval (CI) can be constructed for the difference. If the confidence interval includes zero, the systems are not significantly different [7].

a) Paired observation - Time: In this analysis the confidence interval does not include zero, so we can say that the feature extraction algorithms are different with respect to time. The sign of the difference indicates that the best algorithm was Sampling-8 + Distance: Cosine.

A) Algorithm combined and B) Best single algorithm

- Confiance=80%, CI = [1265.39 , 1292.49]

b) Paired observation - Precision: In this analysis the confidence interval does not include zero, so we can say that the feature extraction algorithms are different with respect to precision. The sign of the difference indicates that the best algorithm was Algorithm Combined + Distance: Cosine.

A) Algorithm combined and B) Best single algorithm

- Confiance=80%, CI = [5.64 ,26.04]

D. One-Factor Experiments

One-Factor Designs are used to compare various alternatives of a single categorical variable [7]. In our case, we compare various distance metrics. A One-factor designs is only valid if the lines do not represent any additional factor [7].

We collect 20 precision samples of the Sampling-8 algorithm using 4 different distance metric (Cosine, Euclidean, Manhattan, Chessboard). And we collect 24 precision samples of the "Combined Approach" algorithm also using the same 4 different metric distance. These 44 observations composed the data from distance metric comparison study. Just an image of each class was used to measure the accuracy using the Cosine distance, Euclidean distance, Manhattan distance and Chessboard distance.

The allocation of variation of the One-Factor Design was:

- Percentage of variation explained by the metric: 12,67%.
- Percentage of experimental error: 87,33%.

We use the ANOVA Table [7] in the analysis of the distance metric comparison study. The results show the F-Value of 1.93 and the F-Table (90-Percentiles) of 2.33 for the Distance Metric. As we can see, the computed F-value is less than F-table value and therefore we conclude that the observed difference in the precision of image retrieval is mostly due to experimental errors and not to any significant difference among the Distance Metric.

E. Experimental Design

Images are very different from each other. There are several pictures of Beach, Dinosaurs, so on. Our initial hypothesis is that images richer in detail may be more difficult to recover with precision. The reason for such a large variability due to experimental error is that the image class should also be considered as a factor, just as we did with the distance metric.

When we took only the first image of each image class, we were saying that time and the precision (the metrics studied) independent of the class. But actually depends. Let's take a look at the data shown in Table I.

TABLE I
EXAMPLE OF WIDE VARIATION IN PRECISION MEASUREMENT IN DIFFERENT IMAGE CLASSES

Class	Relevant Images	Precision (%)
1) Africa	1	**2**
2) Beach	4	8
3) Monuments	5	10
4) Buses	2	4
5) Dinosaurs	50	**100**
6) Elephants	6	12
7) Flowers	45	90
8) Horses	6	12
9) Mountains	16	32
10) Food	1	2

Table I shows the precision measurement collection for the "Sampling-8 + Distance: Cosine" algorithm. As we can see, there is a wide variation between image classes (variation from 2 to 100 - 50 times higher). That was what we call 87% of the experimental error. In other words, it is exactly the intuition we found when we applied the One-Factor Design. The variability between image classes is so great that what we called of experimental error dominates.

In the next subsection (IV-F), we use a 2^k Factorial Design to show the effects of the intra-class variation.

F. 2^k Factorial Design

A 2^k experimental design is used to determine the effect of k factors, each of which have two alternatives or levels. This class of factorial design helps in sorting out factors in the order of impact [7].

We use a 2^2 experimental design, a special case of a 2^k experimental design with k=2, to analyze the impact of two factors (distance metric and image class) on two levels each. The levels of distance metric are Cosine distance (level - 1) and Manhattan distance (level + 1) and the levels of image class are Class 1 (level - 1) and Class 5 (level + 1) (See Table I). We chose classes 1 (Africa) and 5 (Dinosaurs), because the precision measurements using the "Sampling-8 + Cosine Distance" algorithm to ten classes of the WANG dataset showed the Class 1 as the class with the worst average precision and Class 5 as the class with the best average precision.

As result, 2^k Factorial Design shows that the total variation is 2050.69, of which 2047.56 (99,85%) can be attributed to the image class, only 2.40 (0,12%) can be attribuited to distance metric, and only 0.72 (0.04%) can be attributed to interaction.

TABLE II
LEVELS OF DISTANCE METRIC AND LEVELS OF IMAGE CLASS USED IN THE 2^2 EXPERIMENTAL DESIGN

Distance	Class -1 (Africa)	Class +1 (Dinosaurs)
Cosine (-1)	2,8	48,9
Cosine (+1)	5,2	49,6

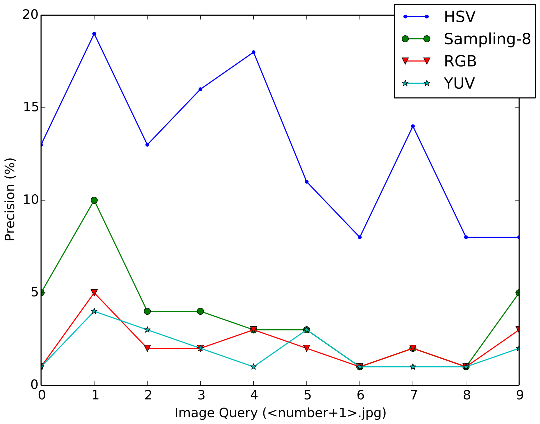

Fig. 3. HSV algorithm was better to recover images in ten images of monuments

With these results we can conclude that the intra-class variation is much more impressive than the distance metric. In other words, more attention must be concentrated in this factor in the CBIR process.

G. Workload Characterization

In order to test multiple alternatives under identical conditions, the workload characterization should be analyzed in detail [7]. To know the variability in a process, we used ten images of the same class and measure the accuracy for each of methods to check whether there was a significant intra-class variation or not. The results are detailed below.

Categorization of HSV Feature Extraction: HSV (hue-saturation-value) algorithm was better to recover monuments pictures (see Figure 3). But HSV seems to be very sensitive to particular image, although it is the best method to monuments. We can see a sudden fall of precision from 18 to 8 (from image 5.jpg to image 7.jpg) in Figure 3. The reason is because the HSV algorithm is sensitive to light. HSV also was the best algorithm to retrieve pictures of “Beach”, “Bus” and “Elephants” of the WANG dataset.

On the other hand, HSV was to bad to recover pictures of dinosaurs, because the images of dinosaurs have a controlled environment and low illumination. HSV also was the worst algorithm to retrieve pictures of “Africa”, “Flowers”, “Foods”, “Horses”.

More Accurate Algorithm: The most accurate algorithm was HSV (40% of the image classes). Sampling-8 was the second best (30%). YUV was accurate in 20% of the image classes. The worst was the RGB (10% of the image classes) - it was unrepresentative (this is what the state of the art also tells).

Final Observations: The type of analysis “The best algorithm for each image type” is the knowledge of the system who says. This is what we of visual pattern recognition (or computer vision) know. Each algorithm will work best in a given situation, environment, etc.

V. CONCLUSION AND FUTURE WORKS

In this paper, we have shown the comparative analysis of the various feature extraction. We also propose an improvement in image retrieval performance by combining these techniques. The results showed that the algorithm proposed in this study (Sampling-8) is faster than the combined approach, but the combined approach is the most accurate approach.

Using a statistical study, we found that only 13% of the variation was explained by the distance metric and 87% was the percentage of the experimental error. A 2^2 Factorial Design showed, for a sample collected, that the variation attributed to the class is approximately 99.85%.

The comparative study showed that each algorithm will work best in a given situation. So, we showed the algorithms that have the best accuracy rate to retrieval images in the WANG dataset and also present the reasons that possibly caused these accuracy rate.

As future work we plan to use new distance metrics and other datasets. We also plan to use the metric of entropy (level of randomness). The idea is calculate the entropy of the entire base, sorting by lower entropy to higher entropy. Finally, use the 2^k Factorial Design (Distance X Entropy) to determine if the Entropy factor has a significant effect or if the observed difference is simply due to random variations caused by measurement error and parameters that were not controlled.

VI. ACKNOWLEDGMENTS

This work was partially financed by Brazilian National Research Council – CNPq (Grant 477457/2013-4 and 449638/2014-6), Minas Gerais Research Foundation – FAPEMIG (Grants APQ-01806-13, APQ-00567-14 and APQ-00768-14), and Coordination for the Improvement of Higher Education Personnel – CAPES.

REFERENCES

[1] D. A. Kumar and J. Esther, “Comparative study on cbir based by color histogram, gabor and wavelet transform,” *International Journal of Computer Applications*, vol. 17, no. 3, pp. 37–44, 2011.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[3] Q. Bao and P. Guo, “Comparative studies on similarity measures for remote sensing image retrieval,” in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 1112–1116.

[4] C. Beecks, M. S. Uysal, and T. Seidl, “A comparative study of similarity measures for content-based multimedia retrieval,” in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1552–1557.

[5] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, “A comparative study of image retargeting,” in *ACM transactions on graphics (TOG)*, vol. 29, no. 6. ACM, 2010, p. 160.

[6] J. Z. Wang, J. Li, and G. Wiederhold, “Simplicity: Semantics-sensitive integrated matching for picture libraries,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 9, pp. 947–963, 2001.

[7] R. K. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*, 1st ed. Wiley, Apr. 1991.