Analytical Comparison of Histogram Distance Measures

Manuel G. Forero(✉), Carlos Arias-Rubio,
and Brigete Tatiana González

Facultad de Ingeniería, Universidad de Ibagué, Ibagué, Colombia
manuel.forero@unibague.edu.co, caar93@hotmail.com,
tatigoq@hotmail.com

Abstract. This paper presents a comparative study of different distance measures used to compare histograms in applications such as pattern recognition, feature selection, image sorting, grouping, identification, indexing, and retrieval. The focus of the study is on how distance measures are affected by variations across images. Different distances between histograms were investigated and tested to compare their performance in retrieving gray scale and color images. A wide range of review papers on calculating distances between histograms was examined. One comparative study was found where histogram bins having zero value were discarded in the calculus of certain distances. We show that this is an inappropriate approach; our tests revealed that zero-value bins should be included to avoid erroneous calculations and achieve a performance advantage over other distance measures.

Keywords: Histogram distances · Bhattacharyya distance · Color distance

1 Introduction

In image analysis, a histogram is a graphical representation of the pixel distribution that describes the amount or frequency of different image intensity values.

When object classification is performed using histograms, the underlying model takes into account only the color of the object and ignores its shape and texture. It is also important to mention that a histogram does not contain spatial information about its corresponding image, i.e., the image cannot be recovered from the histogram and two different images can have the same histograms. Therefore, histograms can be latently identical in two different images, containing different objects but sharing color information. In other words, if there is no spatial or shape information, related objects of different colors may be identified as identical when comparing only the color histograms. Despite the difficulties, solutions like color histogram intersections, indexing constant color, cumulative color histograms, and color distances are used to compare images. While there are drawbacks of using histograms for image indexing and classification, employing real-time color in these tasks has several advantages. One of the advantages is that information is faster to compute compared to other approaches, and it has been shown that color-based methods can be effective in identifying objects of known location and appearance.

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 81–90, 2019.
https://doi.org/10.1007/978-3-030-13469-3_10

There are studies that relate color histogram data with physical properties of objects in an image [1]. These studies have shown that physical properties may represent not only luminescence and color of an object but also image geometry and roughness, all together provide a better estimate of object luminescence and color. Different solutions to the issues associated with comparing color histograms are proposed in the literature, for example, Distance Measure. Among the most utilized measures of distance to calculate the degree of similarity between images are: Euclidean distance, histogram intersection, and quadratic distance. In addition, calculating correlation coefficients is applied. There are many papers discussing distance measures, we found, specifically, two studies in the context of histogram comparison for image analysis tasks. The first paper is a comparative study of histogram distances for object identification by Marín [2], whereas the second paper is itself entitled "On measuring the distance between histograms" by Cha et al. [3]. These two papers served as a basis for our study in which the aim is to compare distance measures for calculating the similarity between histograms for image analysis tasks. We also improve the accuracy of some distance measures when indeterminations were found.

In summary, this paper presents a comparative analysis of some of the most popular techniques for measuring distances between histograms. Modifications to the distances with indeterminations are also proposed. These modified algorithms can be employed in tasks such as pattern recognition, feature selection, image classification, grouping, identification, indexing, and retrieval.

2 Distances

2.1 Distances Between Histograms

Generally, a distance can be defined as a numerical metric that defines the shortest line between two points. The distance between two histograms A and B can be defined as a mathematical function that meets the following conditions:

(a) Non-negativity: $d(A,B) \geq 0$, where $d(A,B) = 0 \leftrightarrow A = B$;
(b) Symmetry: $d(A,B) = d(B,A)$;
(c) Triangular inequality: $d(A,C) \leq d(A,B) + d(B,C)$.

Two types of measures are used to calculate distances between histograms. One is called bin to bin; it compares corresponding bins in each of the two histograms one by one (i.e., the first bin of one histogram with the first bin of another one, and so on). The second type of distance measure is called cross-bin; it focuses on the bins adjacent to the one considered. We used the bin to bin measure, where each histogram bin is treated in an independent way, and distances can be calculated from additions and averages.

The definitions of the six different distance measures employed in this study are introduced below.

Bhattacharyya. Bhattacharyya distance is used to assess equality between two distributions; the response represents the nearest distance between them. The equation for the distance is given by [4] as follows:

$$d(H_1, H_2) = -\ln(BC(H_1, H_2)) \tag{1}$$

$$BC(H_1, H_2) = \sum\nolimits_{I=0}^{N-1} \sqrt{H_1(I) * H_2(I)} \tag{2}$$

where $BC(H1, H2)$ is the Bhattacharyya coefficient for discrete probability distributions and N is the number of bins, usually 256, H_1 and H_2 are the first and the second histograms respectively, and $\overline{H_1}$, and $\overline{H_2}$ represent their means calculated as

$$\overline{H_K} = \frac{1}{N} \sum\nolimits_{J=0}^{N-1} H_k(J) \tag{3}$$

Chi-square. Chi-square distance is a statistical measure that compares observed and expected values for a data set. It is defined by the following expression [5]:

$$d(H_1, H_2) = \sum\nolimits_{I=0}^{N-1} \frac{(H_1(I) - H_2(I))^2}{H_1(I)} \tag{4}$$

Correlation. Correlation is a measure of describing the degree of linear dependence between two histograms. Its value varies between −1 and +1. If the result is zero, it means that there is no linear association between the two histograms being compared. It is calculated as follows [5]:

$$d(H_1, H_2) = \frac{\sum_{I=0}^{N-1} \left(H_1(I) - \overline{H_1}\right)\left(H_2(I) - \overline{H_2}\right)}{\sqrt{\sum_{I=0}^{N-1} \left(H_1(I) - \overline{H_1}\right)^2 * \sum_{I=0}^{N-1} \left(H_2(I) - \overline{H_2}\right)^2}} \tag{5}$$

Intersection. Intersection metric is a measure that considers the intersection of two histograms and tells how many gray levels from the first histogram are present in the second one. The equation is provided below [5]:

$$d(H_1, H_2) = \sum\nolimits_{I=0}^{N-1} \min(H_1(I), H_2(I)) \tag{6}$$

Kullback-Leibler (KL). Kullback-Leibler pseudo-distance is an asymmetrical measure that does not meet the condition (b) of the distance definition introduced earlier. This measure originated from the information theory for handling relative entropy. It is used to measure the average bit number required to identify an event from a set of possibilities, and numerically indicates how two histograms resemble each other. It is defined by the following equation [6]:

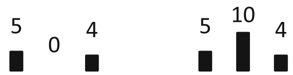
$$d(H_1, H_2) = \sum\nolimits_{I=0}^{N-1} H_1(I)\, log \frac{H_1(I)}{H_2(I)} \tag{7}$$

Euclidian. Euclidean distance is frequently used for evaluating distances in numerical spaces. It is used to determine the bin to bin distance between two histograms and is calculated according to the following equation [3]:

$$d(H_1, H_2) = \sqrt{\sum\nolimits_{I=0}^{N-1} \left(H_1(I) - H_2(I)\right)^2} \tag{8}$$

2.2 Indetermination

As it can be seen from Eqs. (4) and (7), the chi-square and KL distances can be undefined. In his work, Marín [2] simply discard the bins that are zero to avoid this indetermination. However, this solution is inappropriate. For example, according to Marín the two histograms shown in Fig. 1(a) and (b) would be defined as equal [2]. Given that zero bins are discarded making $D_{\text{chi-square}}$ (A, B) and D_{KL} (A, B) to be zero. However, according to the distance definition, distance between two histograms A and B is zero only when A = B. Therefore, in order to solve the indetermination, we considered the following solutions:



(a) (b)

Fig. 1. Example of a critical case when measuring distance between two histograms (a) and (b).

Chi-square. Indetermination can be solved by using Eq. (9) instead of Eq. (4):

$$\sum\nolimits_{I=0}^{N-1} \frac{[H_1(I) - H_2(I)]^2}{H_1(I) + 1} \tag{9}$$

This solution is equal to adding one count to both histograms given that

$$\sum\nolimits_{I=0}^{N-1} \frac{[(H_1(I) + 1 - (H_2(I) + 1)]^2}{H_1(I) + 1} = \sum\nolimits_{I=0}^{N-1} \frac{[(H_1(I) - H_2(I)]^2}{H_1(I) + 1} \tag{10}$$

This solution simply produces a reduction in each one of the addends. Note that a very small value ε must not be added because it would produce a very big addend introducing an error in distance computation. Figures 2(a, b) provide a graphical representation of both the original and the proposed formulas showing how close they are.

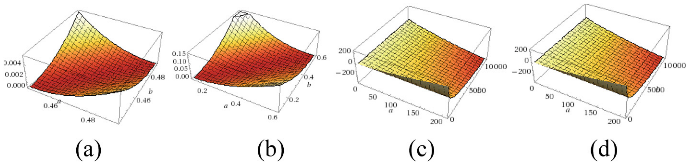


Fig. 2. Graphical representation of chi-square and KL functions: (a, c) original equations, (b, d) modified equations.

KL. Contrary to chi-square, the indetermination in KL distances can be avoided by using the following expression:

$$\sum\nolimits_{I=0}^{N-1} H_1(I) * Log\left(\frac{H_1(I)+\varepsilon}{H_2(I)+\varepsilon}\right) \tag{11}$$

where ε is a small quantity (epsilon). We took $\varepsilon = 0.0001$ for our calculation. We do not take $\varepsilon = 1$ because when $H_1(I)$ and $H_2(I)$ are considerably small, $Log\left(\frac{H_1}{H_2}\right)$ is very different to that of $Log\left(\frac{H_1+1}{H_2+1}\right)$. However, as seen in Figs. 2(c, d), $Log\left(\frac{H_1}{H_2}\right) \cong Log\left(\frac{H_1+\varepsilon}{H_2+\varepsilon}\right)$, when ε is considerably small.

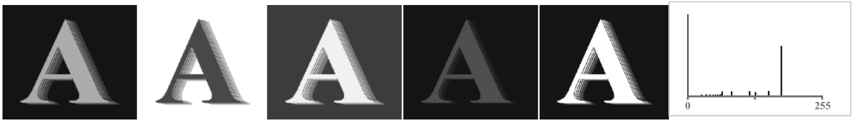
3 Proposal

To test the introduced distances, we considered five synthetic images designed by one of the team members of our University (see Fig. 3), four synthetic histograms were implemented (see Fig. 4), two microscopy images having a background taken with different illumination conditions (see Fig. 5), four microscopy images of a rat brain (acquired from the Instituto de Neurociencias de Castilla y León, Salamanca, Spain, see Fig. 6), and two images of the same objects but with different magnification (see Fig. 7). Distances were calculated for the following cases: between the image in Fig. 3 (a) and each one of its modified variations presented in Figs. 3(b–e); between histograms a, b, c, and d in Fig. 4; and between the images in Figs. 5, 6, and 7. Distances were implemented as plugins for the open source program ImageJ [8].

4 Results

Results are shown in Tables 1, 2, 3, 4 and 5. In column 1 are of the distance between a histogram and itself. It can be seen that, every distance is zero except that for the correlation and intersection. This is because these two measures are not true distances according to the distance definition provided in Sect. 2.1. The correlation distance can even have negative values. Also, the intersection can have a zero value if there are not common bins between two histograms, but it is maximal when they are equal.

In Tables 1, 2, 3, 4 and 5, beside each distance name are indicators in quotes, which correspond to the best equality approximation between histograms, for instance, the best result for Intersection is the largest value, the Correlation distance equals to one, the Chi-square, Bhattacharyya, KL, and Euclidian are all equal to zero. Chi-square distances are very similar between the original and the inverted images than between the original and the high gloss images, which are quite different, since the synthetic image has very few gray levels.



(a) (b) (c) (d) (e) (f)

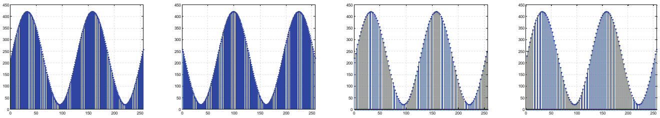
Fig. 3. Synthetic images: (a) original, (b) inverted, (c) high gloss, (d) low gloss, (e) high contrast, (f) histogram of (a).

Table 1. Distances between images in Fig. 3.

Distances	1. Original – Original	2. Original – Inverted	3. Original – High gloss	4. Original – Low gloss	5. Original – Contrast
Chi-square '0'	0.0	9.6771E8	9.676E8	2.834E7	3.4928E7
Intersection '≫'	39831.0	0.0	341.0	30797.0	30702.07
Correlation '1'	1.0	−0.006445	−0.006281	0.9721	0.9669
Bhattacharyya '0'	0.0	1.0	0.9929	0.4386	0.4735
KL '0'	0.0	326107.83	320314.15	64025.39	64340.34
Euclidian '0'	0.0	43992.50	43988.91	7983.16	8013.51

Table 1 shows the distances between synthetic images with a low quantity of gray levels. It can be seen from Table 1 that distances for the original and inverted images are quite dissimilar and there is no correlation between them. Results are appropriate only for the original vs. low gloss and original vs. contrast measures.

The results obtained for synthetic histograms Fig. 4 are shown in Table 2.



(a) (b) (c) (d)

Fig. 4. Synthetic histograms: (a) original (b) inverted, (c) only odd values – "odd hist", (d) only even values – "even hist".

Table 2. Distances between synthetic histograms in Fig. 4.

Distances	1. Original – Original	2. Original – Inverted	3. Original – Odd hist	4. Odd hist – Even hist
Chi-square ‘0’	0.0	296673.98	28084.34	8761770.34
Intersection ‘≫’	56401.0	24474.0	28190.0	0.0
Correlation ‘1’	1.0	−0.9824	0.4725	−0.5525
Bhattacharyya ‘0’	0.0	0.5037	0.5394	1.0
KL ‘0’	0.0	27809.51	181815.72	181671.72
Euclidian ‘0’	0.0	4466.46	2945.744	4164.706

As seen in Table 2, correlation gives a good indication that the histogram in Fig. 4 (b) is the inverted histogram in Fig. 4(a). When comparing “odd hist” with “even hist”, it can be noticed that they are very similar, whereas, the intersection is 0, given that there is no intersection between them. The correlation also does not indicate that these are similar histograms.

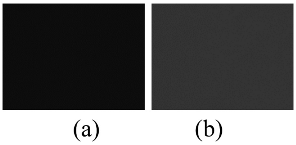


Fig. 5. Microscopy images: (a) intensity 1, (b) intensity 2.

Table 3. Distances between images in Fig. 5.

Distances	1. Intensity 1 – Intensity 1	2. Intensity 1 – Intensity 2
Chi-square ‘0’	0.0	3. 9795E11
Intersection ‘≫’	1428988.0	0.0
Correlation ‘1’	1.0	−0.01538
Bhattacharyya ‘0’	0.0	1.0
KL ‘0’	0.0	1.375019E7
Euclidian ‘0’	0.0	1046609.95

Column 2 in Table 3 shows that the images in Fig. 5 are quite different.

Color images in Fig. 6 were compared in the RGB color space by measuring the distance between each channel histogram and averaging the result as suggested by Prashant [7]. Distances calculated between the images in Fig. 6 are shown in Table 4. As seen in Table 4, the variation in the object intensity can make the distances to indicate that histograms are quite different. This suggests that image intensities must be similar to enable histogram comparison. At the same time, the Bhattacharyya distance is the most robust for intensity variations.

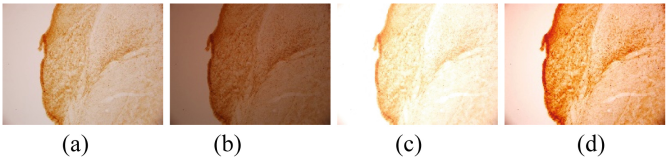


Fig. 6. Images of a rat brain: (a) normal intensity (CRnorm), (b) low light intensity (CRlow), (c) high light intensity (CRhigh), (d) contrast adjustment (CRcontr). (Color figure online)

Table 4. Results of distance measures between the respective images in Fig. 6.

Distances	1. CRnorm – Crnorm	2. CRnorm – CRlow	3. CRnorm – CRhigh	4. CRnorm – CRcontr
Chi-square ‘0’	0.0	1.6388E11	3.8845E13	8.19834E12
Intersection ‘≫’	1.253E7	457896.66	1421693.6	5003024.33
Correlation ‘1’	1.0	−0.2619	−0.0539	0.2354
Bhattacharyya ‘0’	0.0	0.6416	0.32919	0.0
KL ‘0’	0.0	1.115001E8	2.47538E7	8332926.72
Euclidian ‘0’	0.0	2442346.47	5865264.03	2537838.93

Columns 2 and 3 in Table 4 show that brightness variation increase distances. This occurs for all distances except the Bhattacharyya distance in columns 3 and 4, where the most similar values are obtained. To verify this observation, comparisons of RGB color photographs showing objects with similar light intensities at different distances were performed.

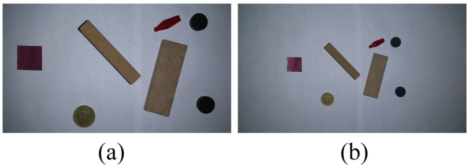


Fig. 7. Images of the same objects: (a) Dist1, (b) Dist2.

Table 5. Distances between images in Fig. 7.

Distances	1. Dist1 – Dist1	2. Dist1 – Dist2
Chi-square ‘0’	0.0	661342.262
Intersection ‘≫’	2073600.0	1603531.0
Correlation ‘1’	1.0	0.8253
Bhattacharyya ‘0’	0.0	0.2647
KL ‘0’	0.0	461700.1
Euclidian ‘0’	0.0	105320.75

The results in Table 5 show that histogram distances are not good indicators of how close two images are when their lightning is quite different. The Bhattacharyya distance shows the best performance even when intensities are dissimilar. When lightning is similar, the correlation and the Bhattacharyya distance show the best results.

The last experiment also shows that other problem in calculating distances occurs if histograms do not contain spatial information about the images and two different images may coincide in their histogram representations.

5 Conclusions

In this work, the performance of six distances measures the similarity between histograms in image analysis tasks was compared. Some of the considered distance measures are not true distances, namely KL distance and correlation.

The chi-square and KL distances were modified to avoid indeterminations when bins are equal to zero. We showed how the proposed solution is more effective compared to the one introduced originally; it prevents two different histograms to appear as equal when indetermination occurs.

It was found that the considered distance measures show bad results when histograms are not continuous, or images of the same objects have a high-intensity variation; only the Bhattacharyya distance showed that two images with the same objects were close when their intensity was very different. When lightning was similar, the Bhattacharyya distance and the correlation performed the best. It was also found that while the correlation is not a true distance, it can be useful for comparing histograms to show how two histograms are related.

In the future, we want to analyze a greater number of distance measures such as the EMD (Earth's moving distance) and test their performance in a higher number of images using histograms from more color spaces.

Acknowledgements. This work was supported by project # 17-461-INT Universidad de Ibagué. We thank Professor Dolores Lopez of the Instituto de Neurociencias de Castilla y León, Universidad de Salamanca for providing the microscopy images.

References

1. Novak, C., Shafer, S.: Method for estimating scene parameters from color histograms. J. Opt. Soc. Am. A **11**(11), 3020–3036 (1994)
2. Marín, P.: Estudio comparativo de medidas de distancia para histogramas en problemas de re-identificación (2015)
3. Cha, S.-H., Srihari, S.: On measuring the distance between histograms. Pattern Recogn. Soc. **35**, 1355–1370 (2002)
4. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943). https://mathscinet.ams.org/mathscinet-getitem?mr=0010358
5. OpenCV Histogram Comparison. http://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/histogram_comparison/histogram_comparison.html. Accessed 30 June 2018

6. Kullback, S.: The Kullback-Leibler distance. Am. Stat. **41**(4), 340–341 (1987)
7. Prashant, I.: Histogram comparison (2013). https://es.scribd.com/doc/168216107/Histogram-Similarity. Accessed 20 May 2018
8. Rasband, W.S., ImageJ. U. S. National Institutes of Health, Bethesda, Maryland, USA (1997–2016). https://imagej.nih.gov/ij/. Accessed 01 Aug 2018