Quantitative Analysis of Similarity Measures of Distributions

Eric Bazan, Petr Dokládal, Eva Dokladalova

Quantitative Analysis of Similarity Measures of Distributions

Eric Bazán[1]
eric.bazan@mines-paristech.fr
Petr Dokládal[1]
petr.dokladal@mines-paristech.fr
Eva Dokládalová[2]
eva.dokladalova@esiee.fr

[1] PSL Research University - MINES ParisTech, CMM - Center for Mathematical Morphology, Mathematics and Systems
35, rue St. Honoré
77305, Fontainebleau Cedex, France

[2] Université Paris-Est, LIGM, UMR 8049, ESIEE Paris
Cité Descartes B.P. 99
93162, Noisy le Grand Cedex, France

Abstract

There are many measures of dissimilarity that, depending on the application, do not always have optimal behavior. In this paper, we present a qualitative analysis of the similarity measures most used in the literature and the Earth Mover's Distance (EMD). The EMD is a metric based on the theory of optimal transport with interesting geometrical properties for the comparison of distributions. However, the use of this measure is limited in comparison with other similarity measures. The main reason was, until recently, the computational complexity. We show the superiority of the EMD through three different experiments. First, analyzing the response of the measures in the simplest of cases; one-dimension synthetic distributions. Second, with two image retrieval systems; using colour and texture features. Finally, using a dimensional reduction technique for a visual representation of the textures. We show that today the EMD is a measure that better reflects the similarity between two distributions.

1 Introduction

In image processing and computer vision, the comparison of distributions is a frequently used technique. Some applications where we use these measures are image retrieval, classification, and matching systems [24]. For these, the distributions could represent low-level features like pixel's intensity level, colour, texture or higher-level features like objects. The comparison could be done using a unique feature, for example, the texture [1, 15], or combining features in a multi-dimensional distribution as the fusion of colour and texture features [18]. In the field of medical imaging, comparing distributions are useful to achieve image registration [25]. More general applications such as object tracking [12, 19] and saliency modeling [5], also use the comparison of distributions. Regarding the number of computer vision applications that make use of the comparison of distributions, the choice of the correct metric to measure the similarity between distributions is crucial.

The Earth Mover's Distance (EMD) [23] is a dissimilarity measure inspired by the optimal transport theory. This measure is considered as true distance because it complies with the constraints of non-negativity, symmetry, identity of indiscernibles, and triangle inequality [20]. The superiority of the EMD over other measures has been demonstrated in several comparative analysis (see for example [22, 23]). Despite this superiority in theory, in practice, this distance continues to be underused for the benefit of other measures. The main reason is the high computational cost due to its iterative optimization process. However, nowadays this should not be a problem thanks to the algorithmic advances to computing efficiently the EMD (see "Notes about EMD computation complexity" in section 4) and the progress of computer processors. Although there are comparative studies (image retrieval scores, for example), in this paper, we illustrate how other similarity measures dramatically fail even on very simple tasks. We use a set of 1D synthetic distributions and two simple image databases (colour and texture-based) to compare a set of similarity measures through two image retrieval systems and a visual representation in low-dimensional spaces. We show that, surprisingly, no metric but the EMD yields to classify and give a coherent visual representation of the images of the databases (see Figs. S2 and S1 in the supplementary material file). In this paper, we want to emphasize the importance of having a true metric to measure the similarity between distributions.

In this article, we present a new qualitative study of some popular similarity measures. Our primary objective is to show that not all measures express the difference between distributions adequately. Also, we show that today the EMD is a competitive measure concerning computing time. Among the similarity measures that we compare are some of the most used bin-to-bin class methods; the histogram intersection and correlation [19], the Bhattacharya distance [25], the χ^2 statistic and, the Kullback-Leibler divergence [12].

This paper is organized as follows: in section 2, we describe and discuss some properties of the bin-to-bin measures and we expose the geometrical properties of the EMD. Then, in section 3, we show the performance of the different similarity measures with a one-dimensional test, with two image classifiers; one based on colour (3D case) and other based on texture (2D case) information and, with a dimensionality reduction using the multidimensional scaling (MDS) technique. Finally, in section 4, we close this work with some reflections about EMD and optimal transport in the field of image processing and computer vision.

2 Similarity Measures Review

Similarity Measures Notation. In many different science fields, there is a substantial number of ways to define the proximity between distributions. In abuse of language, the use of synonyms such as *similarity*, *dissimilarity*, *divergence* and, *distance*, complicates the interpretation of such a measure. Here we recall a coherent notation, used throughout this work.

A ***distance***, from the physical point of view, is defined as a quantitative measurement of how far apart two entities are. Mathematically, a distance is a function $d : M \times M \to \mathbb{R}^+$. We say that d is a ***true distance*** if $\forall (x,y) \in M \times M$ it fulfills the following properties.

1. Non-negativity: $d(x,y) \geq 0$

2. Identity of indiscernibles: $d(x,y) = 0$ if and only if $x = y$

3. Symmetry: $d(x,y) = d(y,x)$

4. Triangle inequality: $d(x,y) \leq d(x,z) + d(z,y)$

From this definition, we can define other distances depending on which properties are (or not) fulfilled. For example, ***pseudo-distances*** do not fulfill the identity of indiscernibles criterion; ***quasi-distances*** do not satisfy the symmetry property; ***semi-distances*** do not fulfill the triangle inequality condition; and ***divergences*** do not comply with the last two criteria [11].

According to the measure, the numerical result could represent the similarity or the dissimilarity between two distributions. The ***similarity*** and the ***dissimilarity*** represent, respectively, how alike or how different two distributions are. Namely, a similarity value is higher when the distributions are more alike while a dissimilarity value is lower when the distributions are more alike. In this paper we use the term ***similarity*** to refer either how similar or how dissimilar two distributions are. If distributions are close, they will have high similarity and if distributions are far, they have low similarity.

2.1 Bin-to-Bin Similarity Measures

In computer vision, distributions describe and summarize different features of an image. These distributions are discretized by dividing their underlying support into consecutive and non-overlapping bins p_i to generate histograms. Let $\mathbf{p}$ be a histogram which represents some data distribution. In the histogram, each bin represents the mass of the distribution that falls into a certain range; the values of the bins are non-negative reals numbers.

The bin-to-bin measures compare only the corresponding bins of two histograms. Namely, to compare the histograms $\mathbf{p} = \{p_i\}$ and $\mathbf{q} = \{q_i\}$, these techniques only measure the difference between the bins that are in the same interval of the feature space, that is, they only compare bins p_i and q_i $\forall i = \{1,\ldots,n\}$, where i is the histogram bin number and n is total number of bins. The Table 1 summarizes the bin-to-bin measures we compare.

The **histogram intersection** [26] is expressed by a *min* function that returns the smallest mass of two input bins (see Eq. 1). As a result, the measure gives the number of samples of $\mathbf{q}$ that have corresponding samples in the $\mathbf{p}$ distribution. According to the notation defined at the beginning of section 2, the histogram intersection is a similarity measure.

The **histogram correlation** gives a single coefficient which indicates the degree of relationship between two variables. Derived from the Pearson's correlation coefficient, this measure is the covariance of the two variables divided by the product of their standard deviations. In Eq. 2, $\overline{\mathbf{p}}$ and $\overline{\mathbf{q}}$ are the histogram means. Since this measure is a pseudo-distance (the resulting coefficient is between -1 and 1), it expresses the similarity of the distributions.

The χ^2 **statistic** measure comes from the Pearson's statistical test for comparing discrete probability distributions. The calculation of this measure is quite straightforward and intuitive. As depicted in Eq. 3, the measure is based on the difference between what is actually observed and what would be expected if there was truly no relationship between the distributions. From a practical point of view, it gives the dissimilarity between two distributions.

The **Bhattacharyya distance** [2] is a pseudo-distance which is closely related to the Bhattacharyya coefficient. This coefficient, represented by $\sum_i \sqrt{p_i q_i}$ in Eq. 4, gives a geometric interpretation as the cosine of the angle between the distributions. We normalize the values of this measure between 0 and 1 to express the dissimilarity between two distributions.

The **Kullback-Leibler divergence** [14] measures the difference between two histograms from the information theory point of view. It gives the relative entropy of $\mathbf{p}$ with respect to $\mathbf{q}$

Histogram Intersection: $$d_{\cap}(\mathbf{p},\mathbf{q}) = 1 - \frac{\sum_i \min(p_i,q_i)}{\sum_i q_i} \quad (1)$$	Histogram Correlation: $$d_C(\mathbf{p},\mathbf{q}) = \frac{\sum_i (p_i-\overline{\mathbf{p}})(q_i-\overline{\mathbf{q}})}{\sqrt{\sum_i (p_i-\overline{\mathbf{p}})^2 \sum_i (q_i-\overline{\mathbf{q}})^2}} \quad (2)$$
χ^2 Statistic: $$d_{\chi^2}(\mathbf{p},\mathbf{q}) = \sum_i \frac{(p_i-q_i)^2}{q_i} \quad (3)$$	Bhattacharyya distance: $$d_B(\mathbf{p},\mathbf{q}) = \sqrt{1 - \frac{1}{\sqrt{\overline{\mathbf{p}}\overline{\mathbf{q}}n^2}} \sum_i \sqrt{p_i q_i}} \quad (4)$$
Kullback-Leibler divergence: $$d_{KL}(\mathbf{p},\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i} \quad (5)$$	Total Variation distance: $$d_{TV}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_i |p_i - q_i| \quad (6)$$

Table 1: Mathematical representation of bin-to-bin measures

(see Eq. 5). Although this measure is one of the most used to compare two distributions, it is not a true metric since it does not fulfill the symmetry and the triangle inequality properties described in section 2.

We can find other measures in the literature that represent the similarity between distributions, for example, the **Lévy–Prokhorov metric** [21] and the **total variation distance** [4]. The first one defines the distance between two probability measures on a metric space with its Borel sigma-algebra. However, the use of this is not very frequent in the area of computer vision because of the implementation complexity [4]. The second one equals the optimal transport distance [7] in the simplified setup when the cost function is $c(x,y) = 1, x \neq y$. For countable sets, it is equal to the L_1 norm. Taking this into account, we only use the first five bin-to-bin measures defined in the following experiments.

2.2 The Earth Mover's Distance

Earth Mover's Distance is the term used in the image processing community for the optimal transport; in other areas, we also find this measure referred to as the Wasserstein distance [9] or the Monge-Kantorovich problem [4, 10]. This concept lays in the study of the transportation theory which aims for the optimal transportation and allocation of resources. The main idea behind the optimal transport is simple and very natural for the comparison of distributions. Let $\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \beta_j \delta_{y_j}$ be two discrete measures supported in $\{x_1,\ldots,x_n\} \in \mathcal{X}$ and $\{y_1,\ldots,y_n\} \in \mathcal{Y}$, where α_i and β_j are the weights of the histograms bins α and β; δ_{x_i} and δ_{y_j} are the Dirac functions at position x and y, respectively. Intuitively, the Dirac function represents a unit of mass which is concentrated at location x. This notation is equivalent to the one proposed in [23] where δ_{x_i} is the central value in bin i while α_i represents the number of samples of the distribution that fall in the interval indexed by i.

The key elements to compute the optimal transport are the cost matrix $\mathbf{C} \in \mathbb{R}_+^{n\times m}$, which define all pairwise costs between points in the discrete measures α and β, and the flow matrix (optimal transport matrix) $\mathbf{F} \in \mathbb{R}_+^{n\times m}$, where f_{ij} describes the amount of mass flowing from bin i (or point x_i) towards bin j (or point x_j). Then the optimal transport problem consists in finding a total flow $\mathbf{F}$ that minimizes the overall cost defined as

$$W_{\mathbf{C}}(\alpha,\beta) = \min\langle \mathbf{C},\mathbf{F}\rangle = \sum_{ij} c_{ij} f_{ij} \quad (7)$$

Placing the optimal transport problem in terms of *suppliers* and *consumers*; for a supplier i, at some location δ_{x_i}, the objective is to supply α_i quantity of goods. On the other hand, a

consumer j, at some location δ_{y_j}, expects to recieve at most β_j quantity of goods. Then, the optimal transport problem is subject to three constraints, $\forall i \in \{1,\dots,n\}$, $j \in \{1,\dots,m\}$

1. Mass transportation (positivity constraint): $f_{ij} \geq 0 : i \to j$.
2. Mass conservation (equality constraint): $\sum_j f_{ij} = \alpha_i$ and $\sum_i f_{ij} = \beta_j$.
3. Optimization constraint: $\sum_{ij} f_{ij} = \min\left(\sum_i \alpha_i, \sum_j \beta_j\right)$.

Then, we define the Earth Mover's Distance as the work $W_{\mathbf{C}}$ normalized by the total flow

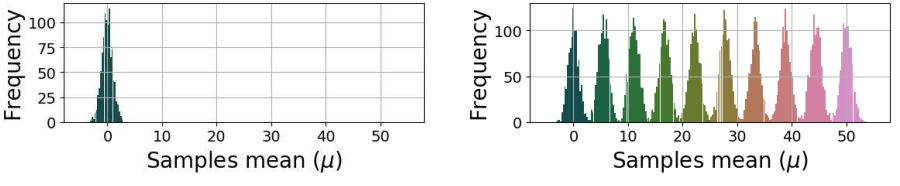
$$d_{EMD}(\alpha,\beta) = \frac{\sum_{ij} c_{ij} f_{ij}}{\sum_{ij} f_{ij}} \tag{8}$$

The importance of the EMD is that it represents the distance between two discrete measures (distributions) in a natural way. Moreover, when we use a *ground distance* as the cost matrix $\mathbf{C}$, the EMD is a true distance. Peyré and Cuturi [20] show the metric properties of the EMD. To show these advantages, we developed a series of experiments described in the next section.

3 Comparative Analysis of Similarity Measures

3.1 One-Dimensional Case Study

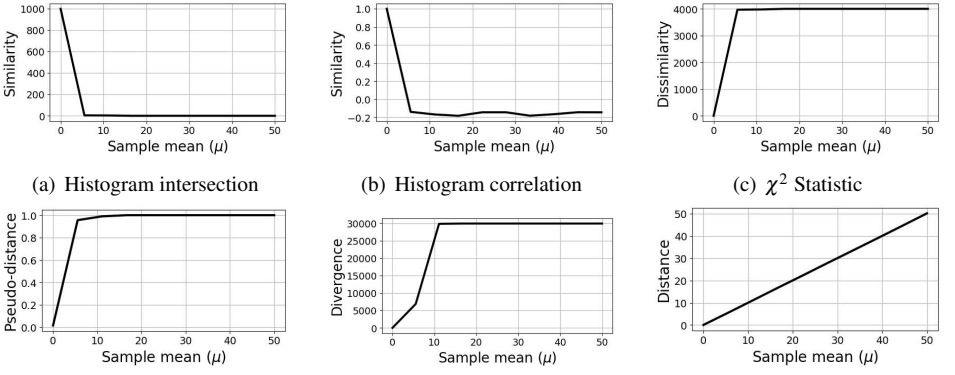
To compare the measures described in section 2 in the simplest scenario, we use a set of one-dimensional synthetic distributions. We create a source distribution and a series of target distributions (see Fig. 1). Both, source and target distributions, has 1000 samples and are random normal distributions. The unique difference between them is that the mean of the target distributions (μ) is increasing five units with respect to the previous distribution.



(a) Source distribution (b) Target distributions

Figure 1: Source and target synthetic distributions

Since the distributions' mean value increases linearly, we expect that the similarity measure has an equivalent response, i.e., that the similarity decreases when the difference of the source and target means increases. In Fig. 2, we can see the response of the bin-to-bin measures and the EMD. Among the bin-to-bin measures, those that give a coefficient of dissimilarity (χ^2 statistic, Bhattacharyya pseudo-distance and K-L divergence) rapidly saturate and stick to a maximum value; while for those that give a coefficient of similarity (histogram correlation and intersection), their value falls rapidly to zero. We can interpret these behaviors as follows. When the bins p_i, q_i do not have any mass in common, the bin-to-bin measures fail in taking into account the mutual distance of the bins. They could consider that



(a) Histogram intersection (b) Histogram correlation (c) χ^2 Statistic

(d) Bhattacharyya (e) Kullback-Leibler (f) EMD

Figure 2: Distances between the source and target distributions

the distributions are precisely at the same distance (there is no difference between them), or that the distributions are entirely dissimilar. The only measure that presents a convenient behavior with the increasing difference of the means is the EMD. This is due to taking into account the *ground distance* $\mathbf{C}$ of the matching bins (see above, Eq. 8). One can argue that for applications like image retrieval finding the most similar distribution is sufficient to find the alike image, or texture, whereas the ordering on the other ones is irrelevant. In the following experiment, we show that this intuition is incorrect and that even in an overly simple case the bin-to-bin measures are not the best choice.

3.2 Image Retrieval Systems

We develop two image retrieval systems as a second comparison test, one based on colour information and the other based on texture information. For the classifiers, we use different databases. The first one contains 24 different colour images of superhero toys[1]. It has 12 classes with two samples per class. The two first rows in Fig. 3 show some examples of the superhero toys and their variations (change of the angle of the toy or the addition of accessories). The second database [16] is composed of images belonging to different surfaces and materials (see the last two rows in Fig. 3). The database contains 28 different classes; it contains different patches per class.

We compare the performance of six out of the eight measures described in section 2 in the following way. First, we divide the database samples into *model images* and *query images*; each class only has one model and one query image. We take an image of the query set and compare its colour/texture distribution (source distribution) with the colour/texture distribution of all model images (target distributions). Then, we order the images from the most similar to the most dissimilar image. We repeat this process for all the images in the query set[2].

Image colour distribution. We use 3D histograms to represent the distribution of colour pixels. Since the superhero images are very simple and do not present significant challenges,

[1] CC superheros images courtesy of Christopher Chong on Flickr

[2] The image classification systems (colour-based and texture-based) and the datasets used for this work are available at https://github.com/CVMethods/image_clasifier.git

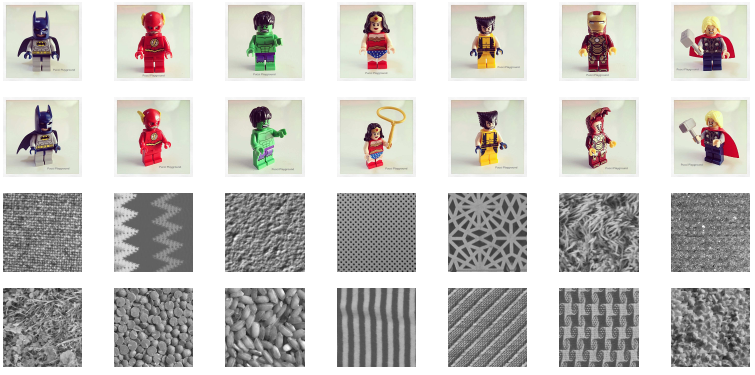


Figure 3: Some samples from the colour and texture databases

i.e., the images possess a very distinctive colour palette and do not present textures or important changes in lighting, any similarity measure should be sufficient to perform the image retrieval. However, image retrieval systems are sensitive to the representation and quantification of the colour image pixels. We show this effect by varying the colour space and the colour quantization level in the image classifier. For the colour space, we represent the images in the RGB, HSL, and LAB colour spaces. For the colour quantization level, we represent the colour space in histograms of 8, 16 and 32 bins per channel.

Image texture distribution. We use a family of Gabor filters as texture descriptors to obtain a distribution (2D histograms) that models the texture of the images. This type of filters models the behavior of the human visual cortex [8], so they are very useful in computer vision applications to represent textures. [17]. The family of Gabor descriptors is represented by the mother wavelet

$$\mathcal{G}(x,y,\omega,\theta) = \frac{\omega^2}{4\pi\kappa^2} e^{-\frac{\omega^2}{8\kappa^2}\left(4\hat{x}^2+\hat{y}^2\right)} \cdot \left[e^{i\kappa\hat{x}} - e^{\frac{\kappa^2}{2}}\right] \tag{9}$$

where $\hat{x} = x\cos\theta + y\sin\theta$ and $\hat{y} = -x\sin\theta + y\cos\theta$. The 2D texture histograms are a function of the energies of the Gabor responses according to the frequency ω, the orientation θ and the constant κ in the Gabor wavelet. For example, a histogram with $\kappa \approx \pi$ and 6 bins means that there are 6 different frequencies to a bandwidth of one octave and 6 different orientations. The energy of a Gabor response is given by,

$$E_{\omega,\theta} = \sum_{x,y} |W_{x,y,\omega,\theta}|^2 \tag{10}$$

where $W_{x,y,\omega,\theta}$ is the response of the convolution of a 2D wavelet with frequency ω and orientation θ with an image.

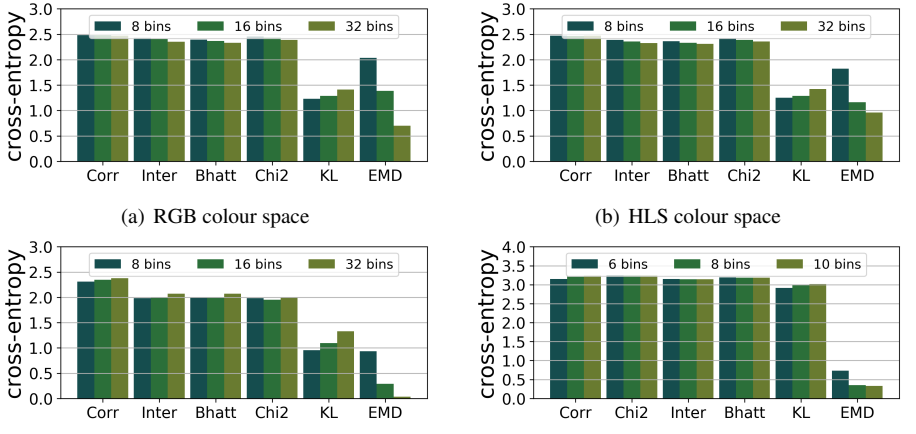
3.2.1 Image Retrieval Systems Evaluation

We create a comparison benchmark for the six similarity measures. First, we normalize the distances given by the different methods between 0 and 1, where a value close to 0 means the distribution the most similar to the source distribution and 1 the most dissimilar. Then we transform the normalized distances into probabilities using a softmax function $SM(\mathbf{d}) = \frac{e^{\mathbf{d}}}{\sum_i e^{d_i}}$. The vector $\mathbf{d} = \{d_i\}, \forall i = \{1,\ldots,m\}$, represents the distances between the query image

to the m images in the database. Considering the softmax function of the distances vector as a classification probability $SM(\mathbf{d}) = \hat{\mathbf{y}}$, we compute the cross-entropy [3] considering the classification ground truth $\mathbf{y}$.

$$H(\mathbf{y},\hat{\mathbf{y}}) = -\sum_i y_i \log \hat{y}_i \tag{11}$$

We can interpret the cross-entropy value as the confidence level of the image classifier for some given metric, feature space (colour or texture) and histogram size. When this value is very close to zero, it indicates a perfect classification of the query image. In Fig. 4, we note the superiority of the EMD over the other measures in both colour and texture-based classifiers.



(a) RGB colour space

(b) HLS colour space

(c) LAB colour space

(d) Texture Gabor energy space (Eq. 10)

Figure 4: Cross entropy value of image retrieval systems (colour and texture) using different similarity measures

With the image retrieval systems, we can highlight interesting aspects of the EMD and the use of bin-to-bin measures in the comparison of distributions. First, we see the importance of the selection of the colour space and the compression level of the feature space (histogram size). The effect of discretization in the bin-to-bin measures is counter-intuitive by the fact that the error increases slightly when the number of bins increases. The explanation could be a poorer intersection of mass distributions. In the case of EMD, increasing the number of bins improves the classification result. Besides, as expected, in the colour-based classifier the calculation of the EMD in the LAB colour space is better than in the HLS or the RGB. This effect is because the LAB colour space models the colour human perception in the Euclidean space, therefore, the *ground distance* between two colours is easily calculated with the L_2 norm. On the other hand, in the texture-based classifier, we see that increase the number of bins beyond 8 bins does not improve the classification considerably. This is because the histograms with 8 frequencies and 8 orientations represent sufficiently well the image textures.

3.3 Texture Projection Quality Evaluation

We use the multidimensional scaling (MDS) technique [13] as the last evaluation test for the similarity measures on the texture dataset. The MDS allows to geometrically represent

n textures by a set of n points $\{x_1,\ldots,x_n\}$ in a reduced Euclidean space $\mathbb{R}^d$, so that the distances between the points $\hat{d}_{ij} = ||x_i - x_j||_2$ correspond as much as possible to the values of dissimilarity d_{ij} between the texture distributions. To evaluate the quality of the projection, we use the stress value proposed in [13].

$$S = \sqrt{\frac{\sum_{ij}(\hat{d}_{ij} - d_{ij})^2}{\sum_{ij}\hat{d}_{ij}^2}} \tag{12}$$

The stress coeficient is a positive value that indicates how well the distances given by the measures are preserved in the new low-dimensional space, i.e. the lower the level of S, the better the representation of texture in a low-dimensional space (2D in our case). The figure 5 shows how the lowest stress is obtained using the EMD. This is because the MDS technique interprets the distances of the entrance towards distances in a space of low dimension. Given that the EMD is the only measure that is a true metric, not only is the stress level low, but the visual projection is in accordance with the frequency ω and orientation θ used in the Gabor filters (Eq. 9) that model the distribution of the textures (see Fig. S8 in supplemental materials file).

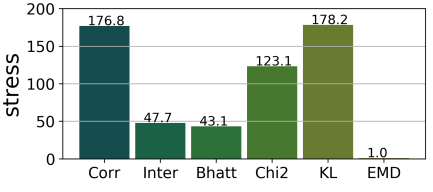


Figure 5: Stress value of the MDS projections using the six principal similarity measures

4 Conclusion

In this work, we compare some of the so-popular bin-to-bin similarity measures with the EMD. We measure their performance in three tests: a one-dimensional analysis with synthetic distributions, with two image classifiers (colour and texture-based) and a visual projection using the MDS technique and the stress as the comparison value. The objective is to show that such measures highly used in the literature to develop complex tasks are not the best choice since they fail even in the most straightforward conditions. We illustrate that the EMD is a true metric [20] that expresses the dissimilarity between distributions naturally.

Results. The experiments of the previous sections show the superiority of the EMD to represent the similarity between distributions. First, the one-dimensional case shows how the bin-to-bin measures saturate (or fall to zero) as soon as the probabilities have empty intersection (see Fig. 1). As for the image retrieval systems, we can see that by correctly choosing the feature image space and a good compression resolution of the distributions (LAB color space with 32 bins in the colour-based system and the Gabor energy with 8 bins in the texture-based system), the EMD performs a perfect classification. However, this is not the case of the other measures because they are not a true distance. Representing the textures in the Euclidean space using the MDS technique, shows another advantage of the EMD. The use of the ground distance $\mathbf{C}$ in the calculation of the optimal transport, making

it possible to transfer 2D texture histograms to a logarithmic-polar space, making the stress value relatively low. To see some examples of the colour-based image retrieval system and the 2D texture projections consult the file with supplementary material.

Notes about EMD computation complexity. We believe that EMD is a depreciated metric only because of excessive calculation time. In the examples developed before, we calculate the EMD using the iterative process of linear programming. Despite this, the calculation is fast enough to develop the image classifier. In comparison with the first EMD algorithm [23], the progress of the computer processors allows to use the same algorithm and be competitive with the bin-to-bin measures. Moreover, a solution to the excessive complexity time and memory consumption are the regularized distances, also called Sinkhorn distances [6]. This entropy-based regularization accelerates the computing time giving a close approximation of the EMD. The regularization of distances allows for creating parallelizable algorithms.

Acknowledgments

This research is partially supported by the Mexican National Council for Science and Technology (CONACYT).

References

[1] Prithaj Banerjee, Ayan Kumar Bhunia, Avirup Bhattacharyya, Partha Pratim Roy, and Subrahmanyam Murala. Local Neighborhood Intensity Pattern–A new texture feature descriptor for image retrieval. *Expert Systems with Applications*, 113:100–115, December 2018. ISSN 0957-4174.

[2] A. Bhattacharyya. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946. ISSN 0036-4452.

[3] Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, 2006. ISBN 978-0-387-31073-2.

[4] Vladimir I. Bogachev and Aleksandr V. Kolesnikov. The Monge-Kantorovich problem: Achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5): 785, 2012. ISSN 0036-0279.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. ISSN 0162-8828.

[6] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2292–2300, USA, 2013. Curran Associates Inc.

[7] Marco Cuturi and David Avis. Ground Metric Learning. *arXiv e-prints*, page arXiv:1110.2306, October 2011.

[8] J. G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, July 1988. ISSN 0096-3518.

[9] Alison L. Gibbs and Francis Edward Su. On Choosing and Bounding Probability Metrics. *Interdisciplinary Science Reviews*, 70:419–435, December 2002.

[10] L. V. Kantorovich. On a Problem of Monge. *Journal of Mathematical Sciences*, 133 (4):1383–1383, March 2006. ISSN 1573-8795.

[11] M. A. Khamsi. Generalized metric spaces: A survey. *Journal of Fixed Point Theory and Applications*, 17(3):455–475, September 2015. ISSN 1661-7746.

[12] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *2011 International Conference on Computer Vision*, pages 2214–2219, November 2011.

[13] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, June 1964. ISSN 1860-0980.

[14] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851, 2168-8990.

[15] R. Kwitt and A. Uhl. Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval. In *2008 15th IEEE International Conference on Image Processing*, pages 933–936, October 2008.

[16] Gustaf Kylberg. The Kylberg Texture Dataset v. 1.0. External Report (Blue Series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, September 2011.

[17] Tai Sing Lee. Image Representation Using 2D Gabor Wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(10):959–971, October 1996. ISSN 0162-8828.

[18] Peizhong Liu, Jing-Ming Guo, Kosin Chamnongthai, and Heri Prasetyo. Fusion of color histogram and LBP-based features for texture image retrieval and classification. *Information Sciences*, 390:95–111, June 2017. ISSN 0020-0255.

[19] S. M. Shahed Nejhum, J. Ho, and Ming-Hsuan Yang. Visual tracking with histograms and articulating blocks. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[20] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *arXiv:1803.00567 [stat]*, March 2018.

[21] Y. Prokhorov. Convergence of Random Processes and Limit Theorems in Probability Theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

[22] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1165–1172 vol.2, September 1999.

[23] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance As a Metric for Image Retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000. ISSN 0920-5691.

[24] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000. ISSN 0162-8828.

[25] Ronald W. K. So and Albert C. S. Chung. A novel learning-based dissimilarity metric for rigid and non-rigid medical image registration by using Bhattacharyya Distances. *Pattern Recognition*, 62:161–174, February 2017. ISSN 0031-3203.

[26] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991. ISSN 1573-1405.

Supplementary Material: Quantitative Analysis of Similarity Measures of Distributions

1 Colour-based Image Retrieval

The figures S1 and S2 present the result of the classification of the images of two superhero toys. The query image is displayed in the upper left. The rows of the image arrangement represent the different measurements used in the retrieval while the columns show the most similar image from left to right in descending order. The numerical values of the distances are below each image; these values are not normalized nor are they on the same scale.

The two examples here show how, under certain disturbances in color distribution, bin-to-bin measurements are not able to identify the correct result. In the case of the wonderwoman toy, the fact that the query image has an extra accessory modifies the color signature of the figure, while in the case of the superman toy the color signature is very close to those toys that contain red and blue colors. These two images were obtained using the LAB colour space and 32 bins for the pixels colour distribution.

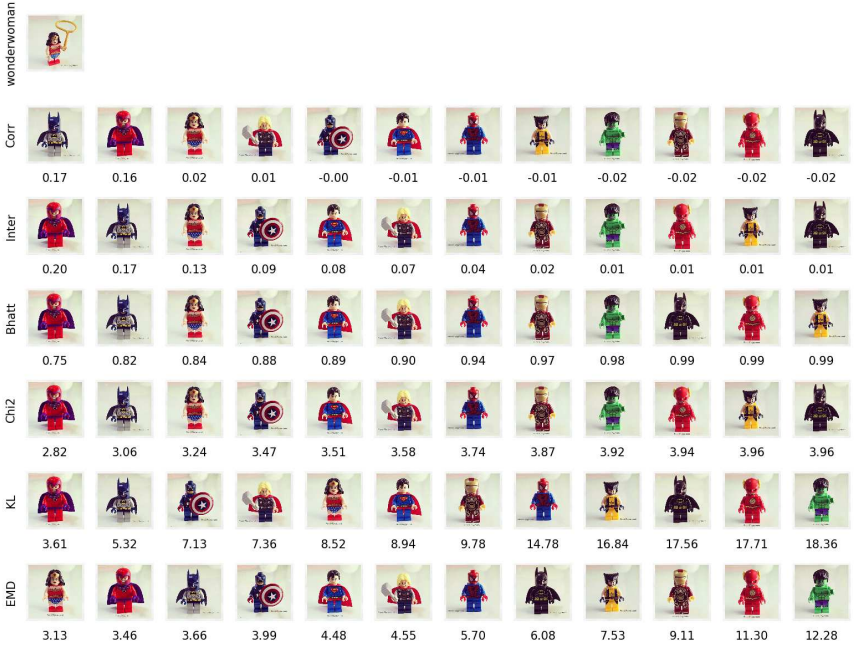


Figure S1: Wonderwoman toy image retrieval example

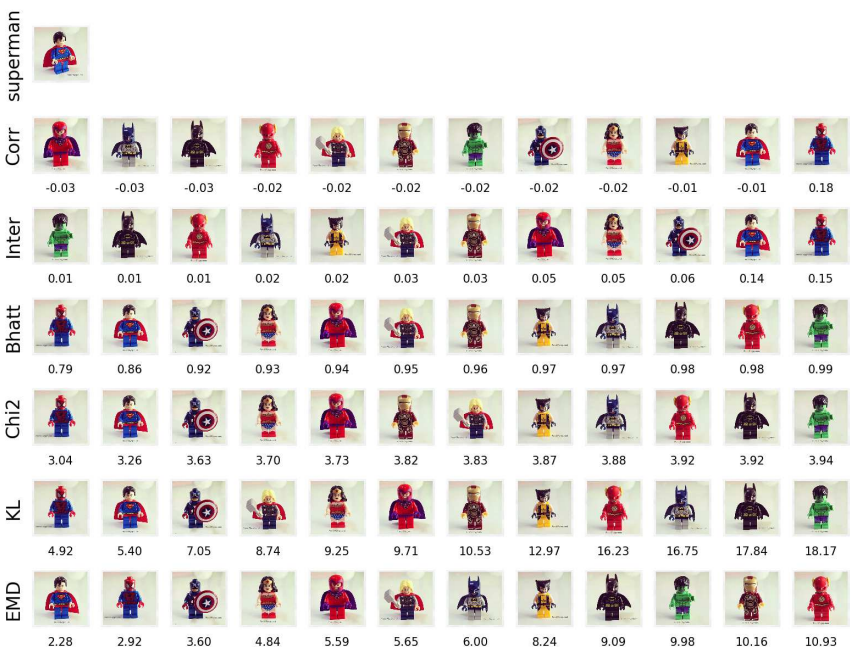


Figure S2: Superman toy image retrieval example

2 Texture Projection Visual Evaluation

The following images serve as a visual tool for the comparative evaluation of the different measures analyzed in the main article. As we described in section 3.3, the MDS technique allows to projecting the textures in a low dimensional space using the distances given by the similarity measures. This representation is carried out in a two-dimensional Euclidean space in our case. In the figures, we can notice that the MDS technique has problems to represent coherently the textures when the input measures are not a metric, i.e., for the bin-to-bin measures. The axis in Figs. S3 to S7 do not correspond with the input space of the textures.

However, in the case of EMD, we can observe that since this measure uses a ground distance to calculate the similarity, we can define the cost matrix $\mathbf{C}_{ij} = c(x_i, y_j)$ of Eq. 7 to be the L_1-distance as

$$c(x_i, y_i) = d((\omega_i, \theta_i), (\omega_j, \theta_j)) = \alpha|\Delta\omega| + |\Delta\theta| \tag{S1}$$

where $|\Delta\omega| = \omega_i - \omega_j$, $\Delta\theta = min(|\theta_i - \theta_j|, \theta_{max} - |\theta_i - \theta_j|)$, and α is a constant that regulates the importence between the orientation and the coarsness of textures. In such a way that it represents the 2D texture distributions into a log-polar space. In the image S8, we can distinguish this effect because the textures are organized concerning their orientation and frequency into a log-polar coordinate space forming a circle. The orientation of texture is represented along the circle while the frequency follows the axis that goes from the outside of the circle to the center, where the lower frequency images remain at the edge of the circle and those with high frequency (and low directionality) are grouped in the center. This behavior is not shown with any of the other measures and is reflected in the stress value of the figure 5.

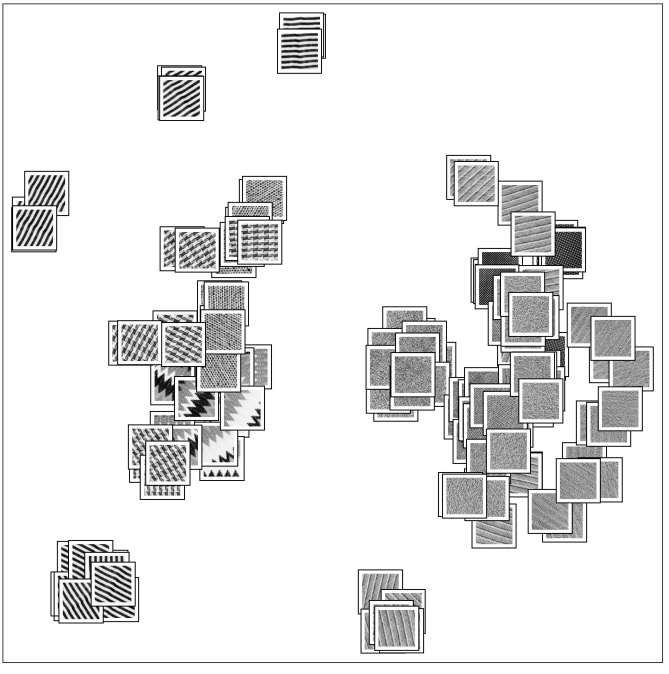


Figure S3: MDS texture projection using the histogram intersection

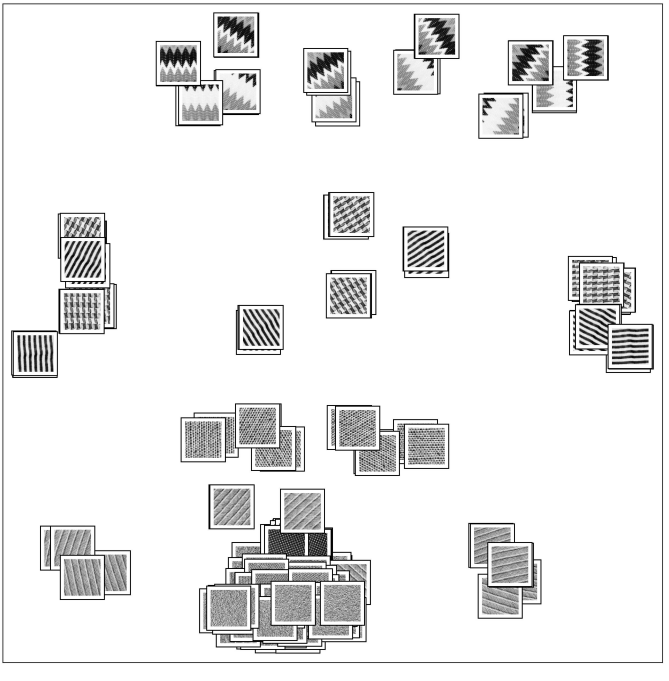


Figure S4: MDS texture projection using the histogram correlation

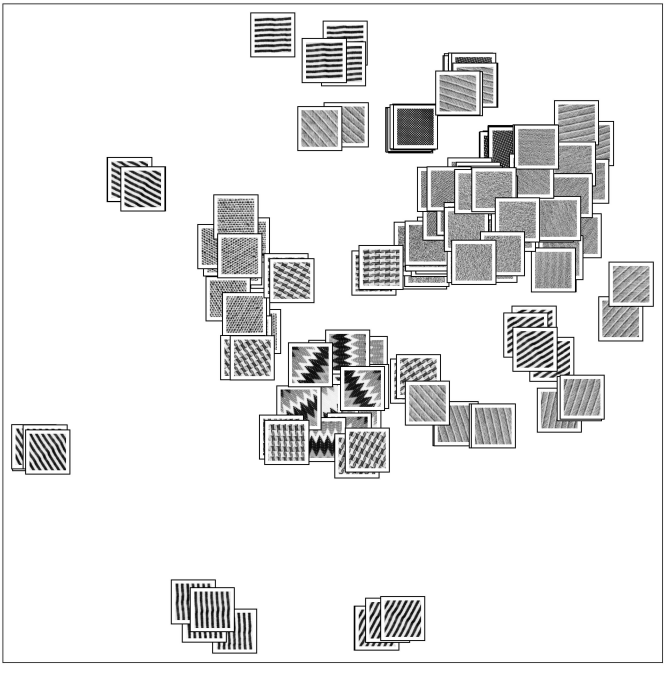


Figure S5: MDS texture projection using the χ^2 statistic

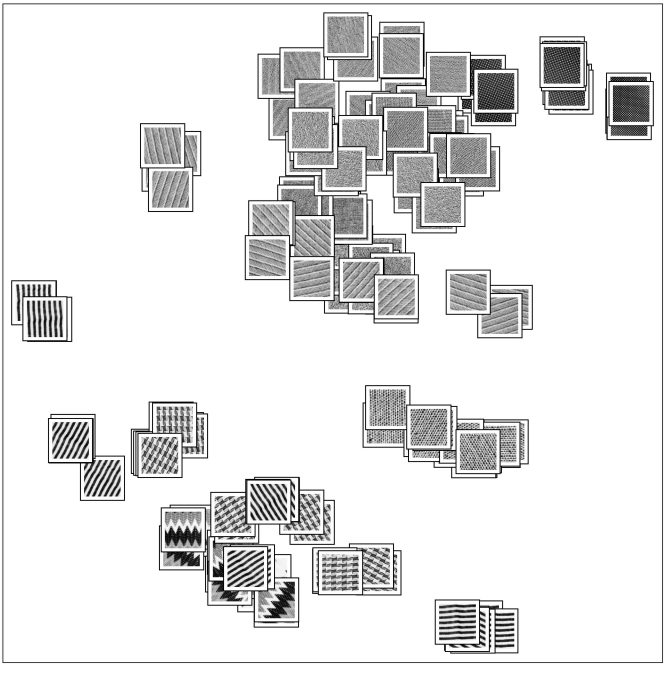


Figure S6: MDS texture projection using the Bhattacharyya distance

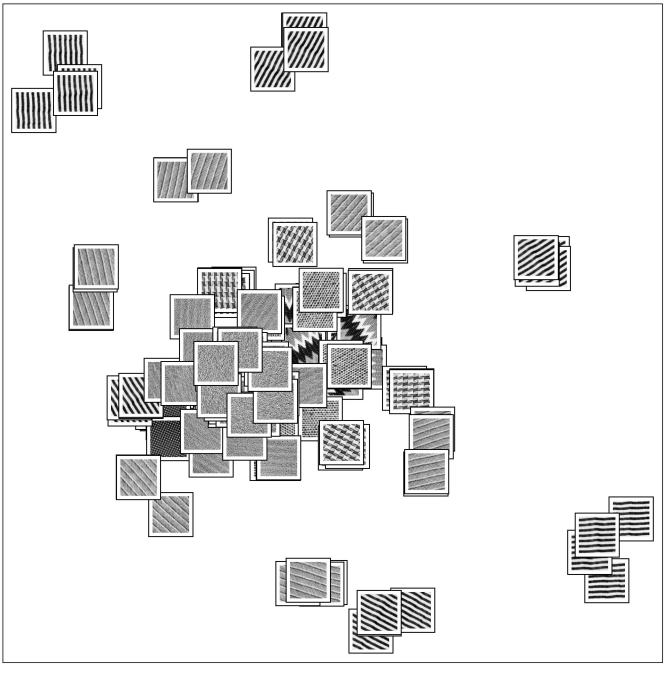


Figure S7: MDS texture projection using the Kullback-Leibler divergence

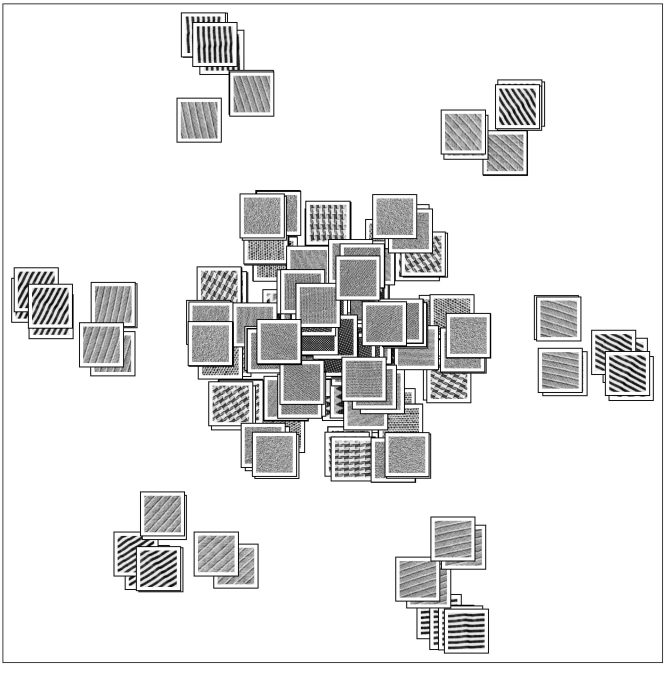


Figure S8: MDS texture projection using the Earth Mover's Distance