

Quantitative Analysis of Similarity Measures of Distributions

Eric Bazán¹, Petr Dokládál¹, and Eva Dokládálová²

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

{eric.bazan, petr.dokladal}@mines-paristech.fr

² Université Paris-Est, LIGM, UMR 8049, ESIEE Paris, Cité Descartes B.P. 99, 93162, Noisy le Grand Cedex, France
eva.dokladalova@esiee.fr

Abstract. The Earth Mover’s Distance (EMD) is a metric based on the theory of optimal transport that has interesting geometrical properties for distributions comparison. However, the use of this measure is limited in comparison with other similarity measures as the Kullback-Leibler divergence. The main reason was, until recently, the computation complexity. In this paper, we present a comparative study of the dissimilarity measures most used in the literature for the comparison of distributions through a color-based image classification system and other simple examples with synthetic data. We show that today the EMD is a computationally efficient measure that better reflects the similarity between two distributions.

Keywords: optimal transport, earth mover’s distance, similarity measures, color image classification, true metric

1 Introduction

The Earth Mover’s Distance (EMD) [14], 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, and triangle inequality [12]. The superiority of the EMD over other measures has been demonstrated in several comparative analysis (see for example [13], [14]). 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. Although there are comparative studies (image retrieval scores, for example), the importance of the errors of the most popular similarity measures has never been illustrated, even for simple tasks. Here, we use a simple database to show that, surprisingly, no metric but the EMD yields the desired result (see below, Fig. 4). In this paper, we want to emphasize the importance of having a true metric to measure the similarity

between distributions. Also, we demonstrate the EMD could be calculated practically with the use of recently created fast algorithms little known in the field of image processing.

In image processing and computer vision, the comparison of distributions is recurrent technique. Some applications where we use these measures are the image retrieval, classification, and matching systems [15]. For these, the distributions could represent low-level features like pixel's intensity level, color, texture, shapes or higher-level features like objects. The comparison could be done using a unique feature (the one-dimensional case for gray-level images comparison) or combining features in a multi-dimensional distribution, for example, the texture [1], [9], or a fusion of color distributions and LBP-based texture features for image retrieval [10]. In the field of medical imaging, distributions comparison are useful to achieve image registration [16]. On the other hand, more general applications such as object tracking [11], [7] and saliency modelling [4] also use the comparison of distributions. Regarding at the number of applications that make use of the comparison of distributions, the choice of the correct metric to measure the similarity between distributions is crucial.

In this article, we present a new comparative study between the EMD and other popular dissimilarity measures. Our primary objective is to show that the compared measures do not express the difference between distributions adequately. Also, we show that today the EMD is a competitive measure concerning computing time. Among the dissimilarity measures we compare are the intersection and correlation of histograms [11], the Bhattacharya distance [16], the χ^2 statistic and the Kullback-Leibler (K-L) divergence [7].

This paper is organized as follows: in the section 2, we describe and discuss some properties of the bin-to-bin measures, while in section 3 we expose the geometrical properties of the EMD. Then, in section 4, we show the performance of the different similarity measures with a one-dimensional case, a color-based image classifier (3D case), and a multivariable case. Finally, in section 5, we close this work with some reflections about EMD and optimal transport in the field of image processing and computer vision.

2 Bin-to-Bin Similarity Measures

In computer vision, the distributions describe and summarize different features of an image. Normally, we compress the image feature distributions dividing their underlying space in a certain number of bins to generate histograms.

Let \mathbf{p} be a histogram which represents some data distribution. In the image domain, these data represent image features such as pixels intensity, color, texture, and others. To construct such histogram, we split the underlying feature space into consecutive and non-overlapping bins p_i . In the histogram, each bin represents the mass of the distribution that falls into its range; the values of the bins are no negative reals numbers.

The bin-to-bin measures metioned in section 1, 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 $q_i \forall i = \{1, \dots, n\}$, where i is the histogram bin number and n is total number of bins. The measures we review here are:

2.1 Histogram Intersection

$$d_{\cap}(\mathbf{p}, \mathbf{q}) = 1 - \frac{\sum_i \min(p_i, q_i)}{\sum_i q_i} \quad (1)$$

Swain and Ballard [17] proposed the histogram intersection algorithm. Mathematically, it is expressed by a *min* function that returns the smallest mass of two input bins. The result of the histogram intersection is the number of samples of \mathbf{q} that have corresponding samples in the \mathbf{p} distribution. We normalized this smallest value by the total number of samples in the corresponding bin q_i . The final score is between 0 and 1, where the highest value means the best match, i.e., the most similar distribution.

2.2 Histogram Correlation

$$d_C(\mathbf{p}, \mathbf{q}) = \frac{\sum_i (p_i - \bar{\mathbf{p}})(q_i - \bar{\mathbf{q}})}{\sqrt{\sum_i (p_i - \bar{\mathbf{p}})^2 \sum_i (q_i - \bar{\mathbf{q}})^2}} \quad (2)$$

The histogram correlation gives a single coefficient that 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, $\bar{\mathbf{p}}$ and $\bar{\mathbf{q}}$ are the histogram means. For this measure, we normalize the resulting coefficient between 0 and 1, where 0 means that the distributions are strongly correlated.

2.3 χ^2 Statistics

$$d_{\chi^2}(\mathbf{p}, \mathbf{q}) = \sum_i \frac{(p_i - q_i)^2}{q_i} \quad (3)$$

This measure comes from Pearson's χ^2 statistical test for comparing discrete probability distributions. It is useful for correspondence analysis because it expresses the fit between observed and expected frequencies.

2.4 Bhattacharyya distance

$$d_B(\mathbf{p}, \mathbf{q}) = \left(1 - \frac{1}{\sqrt{p_i q_i n^2}} \sum_i \sqrt{p_i q_i} \right) \quad (4)$$

The Bhattacharyya distance [2] is a divergence measure which is closely related to the Bhattacharyya coefficient. This coefficient, represented by $\sum_i \sqrt{a_i b_i}$ in Eq. 4, gives a geometric interpretation as the cosine of the angle between the distributions.

2.5 Kullback-Leibler Divergence

$$d_{KL}(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i} \quad (5)$$

The divergence of Kullback and Liebler [8] measures the difference between the two histograms from the information theory point of view. It gives the relative entropy of \mathbf{p} with respect to \mathbf{q} . Although this measure is one of the most used to compare two distributions, it is not symmetrical.

3 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 as the Wasserstein distance or the Monge-Kantorovich problem. 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

$$\boldsymbol{\alpha} = \sum_{i=1}^n \alpha_i \delta_{x_i} \text{ and } \boldsymbol{\beta} = \sum_{j=1}^m \beta_j \delta_{y_j} \quad (6)$$

be two discrete measures supported in $x_1, \dots, x_n \in \mathcal{X}$ and $y_1, \dots, y_m \in \mathcal{Y}$, where α_i and β_j are the weight's bins of histograms $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; δ_{x_i} and δ_{y_j} are the Diracs 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 one the proposed in [14]; δ_{x_i} is the central value in bin i while α_i represents the number of samples of the distribution that falls in the interval indexed by i .

The key elements to compute the optimal transport are the cost matrix $\mathbf{C}_{ij} = c(x_i, y_j)$, which define all pairwise costs between points in the discrete measures $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, 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 aims to find a total flow \mathbf{F} that minimize the overall cost

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

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 receive at most β_j quantity of goods. Then, the optimal transport problem is subject to some constraints, $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$

1. Mass transportation (positivity constraint): $f_{ij} \geq 0 : i \rightarrow j$ allows moving the mass from the suppliers to the consumers.
2. Mass conservation (equality constraint): $\sum_j f_{ij} = \alpha_i$ and $\sum_i f_{ij} = \beta_j$ stays that the suppliers cannot send more goods than they have and the consumers can not receive more goods than they need.
3. Optimization constraint: $\sum_{ij} f_{ij} = \min \left(\sum_i \alpha_i, \sum_j \beta_j \right)$ assures to transport the maximum goods possible from suppliers to consumers.

Then, we define the earth mover's distance as the work $W_{\mathbf{C}}$ normalized by the total flow

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

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 [12] show the metric properties of the EMD. To show these advantages, we developed a series of experiments described in the next section.

4 Comparative Analysis: Bin-to-Bin Measures vs EMD

4.1 One-Dimensional Case: Synthetic data

For the first experiment, we use one-dimensional synthetic distributions to compare the measures described in section 2 in the simplest scenario. We create a source distribution and a series of target distributions. Both source and target distributions are random Gaussian distributions. The unique difference between the source and the target distributions is the increasing mean (μ) of each target histogram. Fig. 1 depicts the histograms of the source and target distributions.

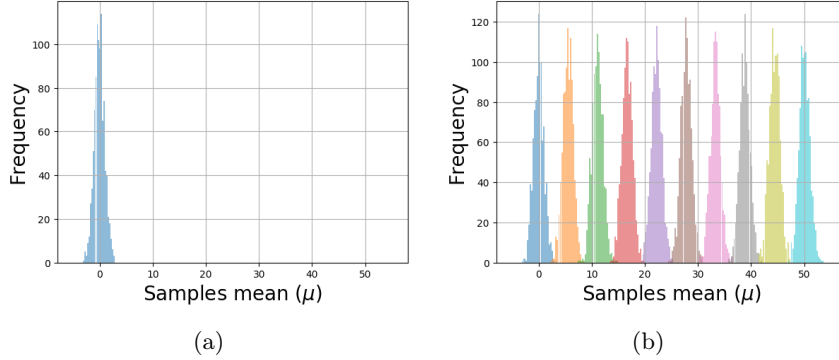


Fig. 1: Synthetic histograms: (a) Source distribution, (b) Target distributions. Each distribution have 1000 samples. The mean of the target distributions is increasing five units with respect to the previous distribution

Since the distributions mean value increases linearly, we expect that the similarity measure also increases linearly. However, in Fig. 2 we can see the response of the bin-to-bin measures and the EMD. For the bin-to-bin measures, the dissimilarity value (χ^2 statistic), the pseudo-distance (Bhattacharyya distance) and the divergence (K-L divergence) rapidly saturate and stick to a maximum value; while the for similarity measures (histogram correlation and intersection), its 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 the distributions are precisely at the same distance (there is no difference between them), or they are entirely dissimilar. The only measure that presents a linear behavior is the EMD. This is due to taking into account the *ground distance* \mathbf{C} of the matching bins (see above, Eq. 8).

4.2 Three-Dimensional Case: Color Image Classification

For the second comparison experiment, we use a database of color images containing 24 photos of different superheroes. The superheroes database has twelve classes with two samples per class. Fig. 3 shows some examples of the database and its variations (change of the angle of the toy or the addition of accessories). The images in the database possess a very distinctive color palette and do not present textures or important changes in lighting. Typically, the comparison of the color distribution should be sufficient to perform an image classification since the images are very simple and do not present significant challenges.

We compare the performance of the six measures described in section 2 in a color-based image classifier. The performance test is simple. First, we divide the database samples into *model images* and *query images*; each class only has one model and one query image. The rows in Fig. 3 show the separation of the

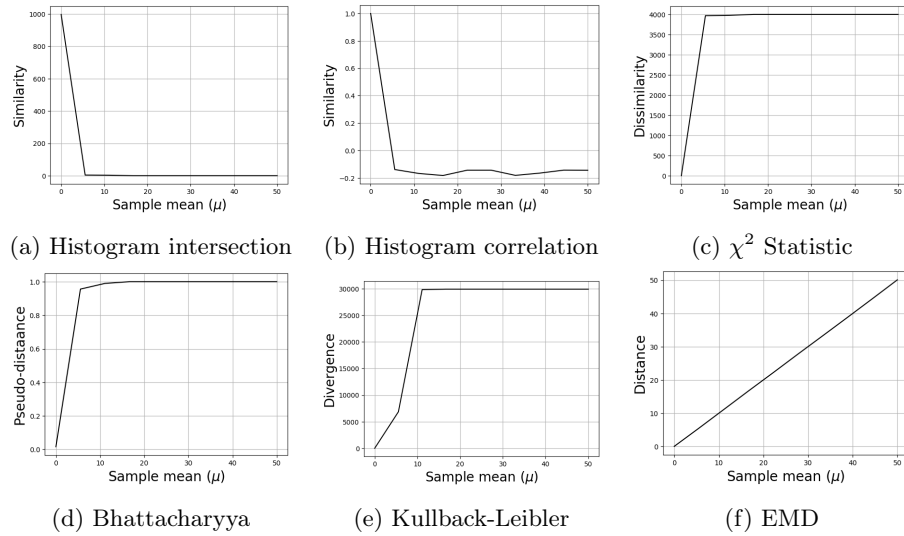


Fig. 2: Distances between the source and target distributions



Fig. 3: Color superheroes database: The columns represent five of the twelve classes of the database. The first row shows the *model images*; the second row shows the *query images*

database. As the second stage, we take an image of the query set and compare its color distribution (source distribution) with the color distribution of all model images (target distributions). Then, we order the distances given by the similarity measure in ascending order, that is, from the most similar to the most dissimilar image. We repeat this process for all the images in the query set.

The image retrieval and classifier systems are sensitive to the representation and quantification of the color image pixels. We show this effect varying the color space and the color quantization level in the image classifier. For the color space, we represent the images in the RGB, HSL, and LAB color spaces. For the color quantization level, we represent the color space in histograms of 8, 16 and

32 bins per channel. Fig. 4 shows the resulting distances for a query image (at the top left of the table) using the LAB color space and histograms of 32 bins per color channel [‡].










































































wonderwoman												
Corr												
	0.17	0.16	0.02	0.01	-0.00	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02
Inter												
	0.20	0.17	0.13	0.09	0.08	0.07	0.04	0.02	0.01	0.01	0.01	0.01
Bhatt												
	0.75	0.82	0.84	0.88	0.89	0.90	0.94	0.97	0.98	0.99	0.99	0.99
Chi2												
	2.82	3.06	3.24	3.47	3.51	3.58	3.74	3.87	3.92	3.94	3.96	3.96
KL												
	3.61	5.32	7.13	7.36	8.52	8.94	9.78	14.78	16.84	17.56	17.71	18.36
EMD												
	3.13	3.46	3.66	3.99	4.48	4.55	5.70	6.08	7.53	9.11	11.30	12.28

Fig. 4: Table of similarity values given for a *query image* representing the images in the LAB color space with histograms of 32 bins. We sort the distances in ascending order, being the left image the most similar.

We create a comparison benchmark for the six dissimilarity measures. First, we normalize the distances given by the different methods between 0 and 1, where 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.

$$S(\mathbf{d}) = \frac{e^{\mathbf{d}}}{\sum_i e^{d_i}} \quad (9)$$

where the $\mathbf{d} = d_i, \forall i = \{1, \dots, m\}$, represents the distances between the query image to the m images in the database.

[‡]The image classification system and an extended superheroes database with 25 superheroes classes are available at https://github.com/E1rc/image_clasifier.git

Finally, considering the softmax function of the distances vector as a classification probability $S(\mathbf{d}) = \hat{\mathbf{y}}$, we compute the cross-entropy [3] considering the classification ground truth \mathbf{y} .

$$H(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i y_i \log \hat{y}_i \quad (10)$$

The cross-entropy is a value that we can interpret as the confidence level of the image classifier concerning the variables we use, i.e., metric, color space and histogram size. When this value is very close to zero, it indicates a perfect classification of the query image. In this experiment, we can highlight interesting aspects of the EMD and the use of bin-to-bin measures in the comparison of distributions. First, we note the superiority of the EMD over the other measures. Besides, we see the importance of the selection of the color space and the compression level of the feature space (histogram size). In the case of EMD, increasing the number of bins improves the classification result, while for other measures this effect is not always evident. On the other hand, as expected, the calculation of the EMD in the LAB color space is better than in the HLS or the RGB. This effect is because the LAB color space models the color human perception in the Euclidean space, therefore, the *ground distance* between two colors is easily calculated with the L_2 norm. In Fig. 5 reflects the facts described below and the sensitivity of the bin-to-bin measures to the different color spaces and histogram size.

4.3 Multidimensional Case: Time and Memory performance

With the experiments of the previous sections, we show the superiority of the EMD to represent the similarity between distributions. However, a clear drawback of this metric is the computation complexity. Regarding Eq. 7, we can see that the EMD is an iterative optimization process depending on the number of bins n of the histograms. The complexity to calculate the distance between a pair of histograms is at least $O(n^3 \log(n))$ [5]. In the area of image processing, this fact is crucial in applications with large scale databases or with high dimensional feature distributions. There, the number of bins is closely related to the number of dimensions of the feature space. The more dimensions the distribution has, the higher the number of bins and the memory consumption to measure the similarity between distributions. One strategy to accelerate the process is to compress as much as possible the data distributions. However, as we can see in Fig. 5, the more we reduce the histogram number of bins, the more precision we lose. Therefore, it leads to a misrepresentation of the similarity between to distributions.

A solution to the excessive complexity time and memory consumption are the regularized distances, also called Sinkhorn distances [5]. This entropy-based regularization accelerates the computing time giving a close approximation of the EMD. To show the complexity of the EMD and the Sinkhorn distances we do a small test. We create two synthetic distributions of n bins. Then, we vary

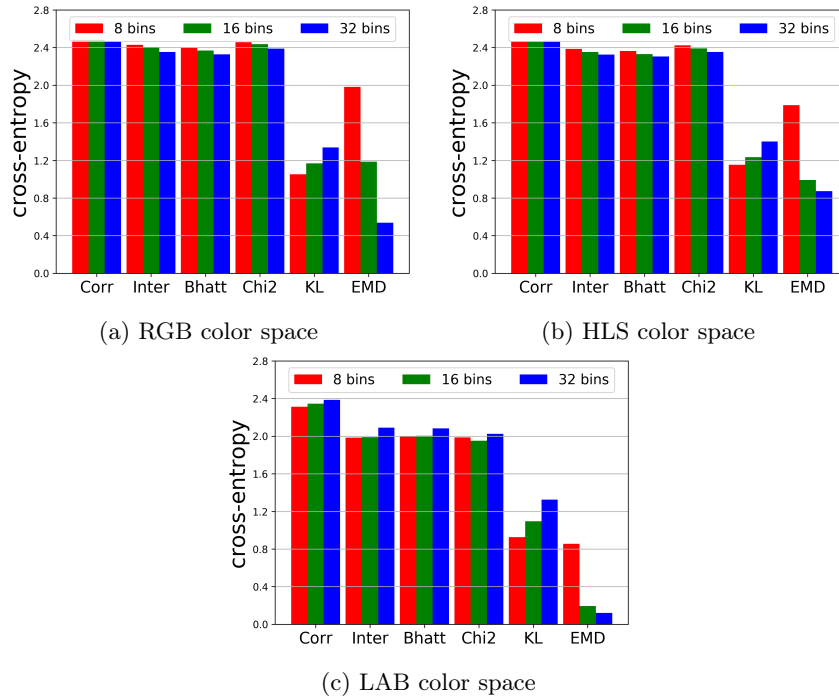


Fig. 5: Cross-entropy values of the image classifier using the six different measures in three color spaces: RGB, HLS, and LAB. The color bars mean the histogram size; red for 8 bins, green for 16 bins and blue for 32 bins per color channel

the number of bins of the histograms and compute both distances. Fig. 6 shows the average time to compute the EMD and the Sinkhorn distances.

The regularization of distances allows creating parallelizable algorithms. It is therefore amenable to large scale executions on parallel platforms such as GPGPUs. In the examples developed in this article, 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 which was limited to 512 bins (see [14]), the progress of the computer processors allows to use the same algorithm and be competitive with the bin-to-bin measures concerning computing time.

5 Conclusion

In this work, we use some of the so-popular bin-to-bin similarity measures and compare them with the EMD. We compare the performance with three simple cases: a one-dimensional case simulating the data distribution, with a very simple color-based image classifier, and with a multidimensional case to see the computation time. The objective is to show that measures highly used in the literature

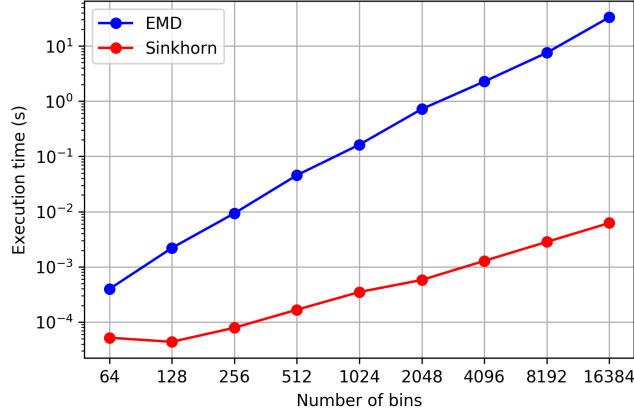


Fig. 6: Computation times of the EMD and the Sinkhorn distances for several histogram sizes. The number of bins in the x axis is the total of bins of the distribution

to develop more complex task (image retrieval systems, image registration, object tracking, saliency modelling, and others) are not the best choice since they fail even in the most straightforward conditions. We show that the EMD true metric that expresses the dissimilarity between distributions naturally.

We believe that EMD is a depreciated metric only because of excessive calculation time. However, after the evolution of computer processors, despite being an iterative optimization process, today their calculation is very fast. Besides, recent works show new strategies such as parallel programming and the use of GPUs [6] using regularization terms [5] that allow to accelerate the calculation time and work with large distributions in dimensions and samples.

Acknowledgments

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

References

1. Banerjee, P., Bhunia, A.K., Bhattacharyya, A., Roy, P.P., Murala, S.: Local Neighborhood Intensity Pattern—A new texture feature descriptor for image retrieval. *Expert Systems with Applications* **113**, 100–115 (Dec 2018)
2. Bhattacharyya, A.: On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **7**(4), 401–406 (1946)
3. Bishop, C.: *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer-Verlag, New York (2006)

4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–1 (2018)
5. Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances (Jun 2013)
6. Flamary, R., Courty, N.: POT Python Optimal Transport library. <https://github.com/rflamary/POT> (2017)
7. Klein, D.A., Frintrap, S.: Center-surround divergence of feature statistics for salient object detection. In: 2011 International Conference on Computer Vision. pp. 2214–2219 (Nov 2011)
8. Kullback, S., Leibler, R.A.: On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79–86 (Mar 1951)
9. Kwitt, R., Uhl, A.: Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval. In: 2008 15th IEEE International Conference on Image Processing. pp. 933–936 (Oct 2008)
10. Liu, P., Guo, J.M., Chamnongthai, K., Prasetyo, H.: Fusion of color histogram and LBP-based features for texture image retrieval and classification. *Information Sciences* **390**, 95–111 (Jun 2017)
11. Nejhum, S.M.S., Ho, J., Yang, M.H.: Visual tracking with histograms and articulating blocks. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (Jun 2008)
12. Peyré, G., Cuturi, M.: Computational Optimal Transport. *arXiv:1803.00567 [stat]* (Mar 2018)
13. Puzicha, J., Buhmann, J.M., Rubner, Y., Tomasi, C.: Empirical evaluation of dissimilarity measures for color and texture. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. vol. 2, pp. 1165–1172 vol.2 (Sep 1999)
14. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover’s Distance As a Metric for Image Retrieval. *Int. J. Comput. Vision* **40**(2), 99–121 (Nov 2000)
15. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 1349–1380 (Dec 2000)
16. So, R.W.K., Chung, A.C.S.: A novel learning-based dissimilarity metric for rigid and non-rigid medical image registration by using Bhattacharyya Distances. *Pattern Recognition* **62**, 161–174 (Feb 2017)
17. Swain, M.J., Ballard, D.H.: Color indexing. *International Journal of Computer Vision* **7**(1), 11–32 (Nov 1991)