A New Image Dissimilarity Measure Incorporating Human Perception

Hamid Shojanazeri, Shyh Wei Teng , Sunil Aryal, Dengsheng Zhang, Guojun Lu
School of Engineering and Information Technology, Faculty of Science and Technology
Federation Univeristy Australia
Churchill, Victoria, Australia
{h.shojanazeri, shyh.wei.teng, sunil.aryal, dengsheng.zhang, guojun.lu}@federation.edu.au

***Abstract*— Pairwise (dis)similarity measure of data objects is central to many applications of image analytics, such as image retrieval and classification. Geometric distance, particularly Euclidean distance ($\ell_2 - norm$), is a popular choice of dissimilarity measure. Because the distance between two data instances is solely based in their spatial positions, the distribution of data around them has no influence in the dissimilarity of two instances. Recently, a data-dependent dissimilarity measure called $m_p - dissimilarity$ is introduced where the dissimilarity of two instances is based on the distribution of data between the two instances. Though it takes the data distribution into account, it completely ignores the geometric proximity of the two data points.**

In this paper, we investigate the strengths and weaknesses of geometric and data-dependent dissimilarity measures. They have their own strengths and weaknesses. To overcome their limitations, we introduce a hybrid dissimilarity measure where the dissimilarity of two instances is based on a combination of the geometric distance and data distribution between them. It uses the strengths of both measures. We present two variants of our proposed hybrid dissimilarity measure. Our empirical results in the image retrieval show that both variants of hybrid dissimilarity measures produce better retrieval results than Euclidean distance and $m_p - dissimilarity$ by themselves.
Keywords—Image retrieval; Dissimilarity measure; Data dependent dissimilarity measure

I. INTRODUCTION

Pairwise (dis)similarity measure of data objects is central to many applications in image analytics, such as image retrieval and classification. For example, in image retrieval, .images are defined by a fixed set of features such as colour, texture, shape and gradient [1-4] A dissimilarity measure is used to rank database images based on their dissimilarity to a given query image. The performances of image retrieval algorithms rely on the effectiveness of the dissimilarity measure used to calculate the similarity of data images to the query.

Miknowski_distance (also known as $\ell_p - norm$), specifically $\ell_2 - norm$ commonly known as Euclidean distance (ED), is a common dissimilarity measure [5-9]. However, $\ell_2 - norm$ only focuses on geometric distance between two vectors and completely ignores the influence of data distribution on the dissimilarity measure. Based on a psychological theory, perceptual dissimilarity between two instances is influenced by the distribution of data, i.e. two instances in a relatively dense area are perceived less similar than in less dense (sparser) area [5, 9, 10]. For example, two red apples among green apples perceptually look more similar than the same two red apples among other red apples. Based on this theory, a mass-based dissimilarity measure, called $m_p - dissimilarity$ (we refer to $\ell_p - norm$ and $m_p - dissimilarity$ by ℓ_p and m_p respectively hereon) has been proposed in [9]. It focuses on the data distribution instead of geometric distance and calculates dissimilarity of two instances based on data density/ distribution in the region covering two instances Two instances in a dense region are more dissimilar than two instances of the same geometric distance located in a sparse region. m_p uses data distribution only [5, 9] and completely ignores the geometric distance that results in some undesirable consequences. For example, it may result in two instances having large geometric distance to be similar if they are in sparse region, i.e., data mass between them is low. However if the same two instances are located in a dense region then m_p will find them dissimilar. So considering only data distribution changes the dissimilarity judgment to a great extent that may be in contrast with the dissimilarity measured by geometric distance. The dissimilarity measured by geometric distance intuitively corresponds to the dissimilarity defined in real three-dimensional world [5]. So to measure the dissimilarity properly m_p and geometric distance should be consistent.

In this paper, we propose a new hybrid dissimilarity measure that incorporates both geometric distance and region density between two data instances. We propose two variants of hybrid dissimilarity measure. In the first variant, the base measure is m_p but it is adjusted by ED in situations where it would fail to express their actual perceived similarity. In the second variant, geometric distance of two instances in each dimension is adjusted by data mass between them. The base measure is distance but it weighted by the mass. The proposed dissimilarity measure can be applied in different applications such as information retrieval and classification. In this paper, we evaluate their performance in image retrieval task.

The rest of the paper is organised as follows. In Section II, we discuss the limitations of ℓ_p and m_p and identify situations where they would fail to express the actual perceived similarity between two instances. The concept of hybrid dissimilarity and its two variants are discussed in Section III. Section IV discusses experimental setup and results. Section VI presents the conclusions and future work.

II. LIMITATIONS OF EXISTING GEOMETRIC AND MASS- BASED DISSIMILARITIES

In this section, we will discuss the limitations with geometric distances and the existing mass-based dissimilarity.

A. Limitation of Geometric Distance

A wide range of geometric dissimilarity measures are discussed in [11]. For the rest of this paper, we will use geometric distance and distance interchangeably. [12, 13] have provided a comprehensive analysis and comparison of the dissimilarity measures in image retrieval. The study in [12] has compared the performance of Histogram Intersection, Minkowski-type Quadratic and Mahalanobis distances. Its results have shown that ED has achieved the best retrieval results.In [13] authors compared the performance of sum of squared of absolute differences, sum of absolute difference, maximum value, Canberra, city block, Minkowski (p=3) and ED. Their results have also confirmed the suitability of ED for image retrieval. An image x is represented by a d–dimensional vector $< x_1, x_2, \cdots, x_d >$ where x_i represent the value of i^{th}feature of x. Generally, ℓ_p between two vectors of x and y is defined as follows [1]:

$$\ell_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^{d} abs(x_i - y_i)^p\right)^{1/p} \quad (1)$$

where $p > 0$, $\|.\|_p$ is the p order norm of a vector, , and $abs(.)$ is the absolute value. $\ell_p - norm$ is a popular choice of distance function as it intuitively corresponds to the distance defined in the real three-dimensional world. It has been widely used as the dissimilarity measure to compare the feature vectors derived from images in many image retrieval systems [6-8, 14, 15].

However, ℓ_p has its limitations. It measures dissimilarity by combining the distance of two data instances in each dimension of the feature space. It computes the dissimilarity between two instances solely based on geometric positions of them in feature space and completely ignores the distribution of data (position of other instances). However, the distribution of image features influences considerably on the perceived dissimilarity between two instances as explained in previous section. Also ℓ_p is under the influence of dominant dimension when combining the distances in each dimension. This may results in situation that dimensions with smaller distances do not contribute proportionally in the calculation of dissimilarity between two instances.

B. Limitation of Mass-based Dissimilarity

To address the discussed limitation of ℓ_p, a data dependent dissimilarity measure has been proposed [9]. This measure is called m_p and it focuses on the data distribution of the dataset instead of simply measuring the distance. It has been shown to perform comparably to or better than ℓ_p in context of information classification and retrieval problems on text, image, music, digits and artificial datasets [5].

m_p is developed based on a distance-density model proposed by Krumhausl [10] and a psychological argument which prescribes that two instances in a sparse region are perceptually more similar than in a dense region. In this measure, the dissimilarity between two instances, x and y, is measured by considering data distribution. It defines a region, $R(x, y)$, between two instances (that encloses the two instances) and finds the data mass. Data mass is the number of data instances from dataset that fall in this region. In order to measure the dissimilarity between two images, x and y, m_p considers the relative positions of x and y with respect to the distribution of the rest of the data in each dimension of their feature vectors and is defined as:

$$m_p(x, y) = \left(\frac{1}{d}\sum_{i=1}^{d}\left(\frac{|R_i(x, y)|}{N}\right)^p\right)^{1/p} \quad (2)$$

where $|R_i(x,y)|$ is the data mass in region of $R_i(x,y)$, and N is the total number of instances in the dataset. The enclosing region is defined as follows. $R_i(x,y) = [\min(x_i,y_i) - \sigma, \max(x_i,y_i) + \sigma\ ]$, σ is a small number and $\sigma \geq 0$.

Data distribution has an effect on the perceived similarity as considered in m_p. However, the geometric distance between two instances should not be ignored, as it intuitively corresponds to the defined dissimilarity in the real three-dimensional world, specifically when the magnitude of vectors in feature space matters. m_p calculates the dissimilarity between two instances solely based on data distribution in the region covering the two instances. If we change the location of the same two instances to a region with different data distribution the dissimilarity may change to the extend that would be in contrast with geometric distance between them which is not desirable.

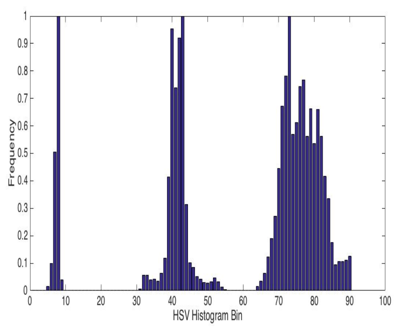
If mass-based dissimilarity and geometric distance measure dissimilarity between two instances properly they must be consistent. m_p considers the lower data mass between two instances as lower dissimilarity and vice versa. However, sometimes the data mass is not consistent with geometric distance. In some situations, m_p will find two instances similar due to low data mass between them while they have a large geometric distance. In other situations, m_p may find two instances dissimilar based on high data mass between them while they are perceptually similar and have small geometric distance. In such situations, mass-based dissimilarity and geometric distance are not consistent and m_p may not retrieve accurate results. A skewed data distribution may result in the discussed situations where m_p would fail to express the actual perceived similarity. In such a data distribution, data mass in some regions is extremely high while it is very low in some other regions. m_p combines the mass-based dissimilarity in each dimension of feature vector to estimate the final dissimilarity. To further illustrate the aforementioned limitation, we provide an example of m_p dealing with skewed distribution in one dimension, as follows.

Figure 1 shows three images and their color histograms from ebay dataset, in which ground truth is based on colour and images of the same color are considered as relevant. Figure 2 shows the distribution of HSV color histograms of ebay dataset in Dimension 6, which is highly skewed. In Figure 1 values in HSV colour histograms for the green shoe, green pot and white shoe in Dimension 6 are 0.1,0.03 and 1. As shown in Figure 2, data distribution in Dimension 6 is highly skewed which means some regions are very dense and some sparse. Data distribution in Dimension 6 shows the region between 0 and 0.1 is very dense while its very sparse between 0.1 and 1. So the number of data fall in the region between 0.1 and 0.03, belong to values of green pot and shoe in this dimension is around 400, that is much higher compared to this number in the region between green and white shoe, 0.1 and 1 which is only 144. However, the distance between the green shoe and green pot is 0.07, which is much smaller compared to 0.9 between green and white shoe .In this situation m_p will find the white shoe more similar to the green shoe in Dimension 6 due to lower data mass between them compared to green pot. So the highly skewed data distribution has resulted in the situation that dissimilarity measured by m_p is not consistent with distance, which reflects the perceptual dissimilarity more accurately.

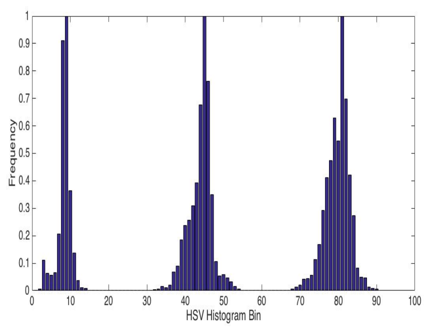
(a) Green Shoe

(b) Green Pot

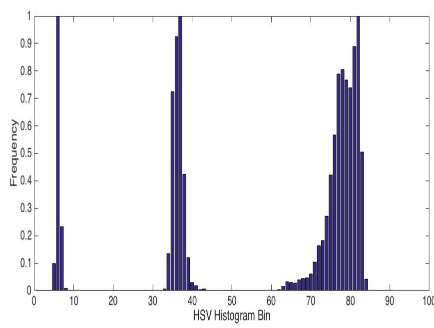
(c) White Shoe



(d) HSV Histogram of Green Shoe



(e) HSV Histogram of Green Pot



(f) HSV Histogram of White Shoe

Figure 1. Sample of images from ebay dataset and their colour histograms.

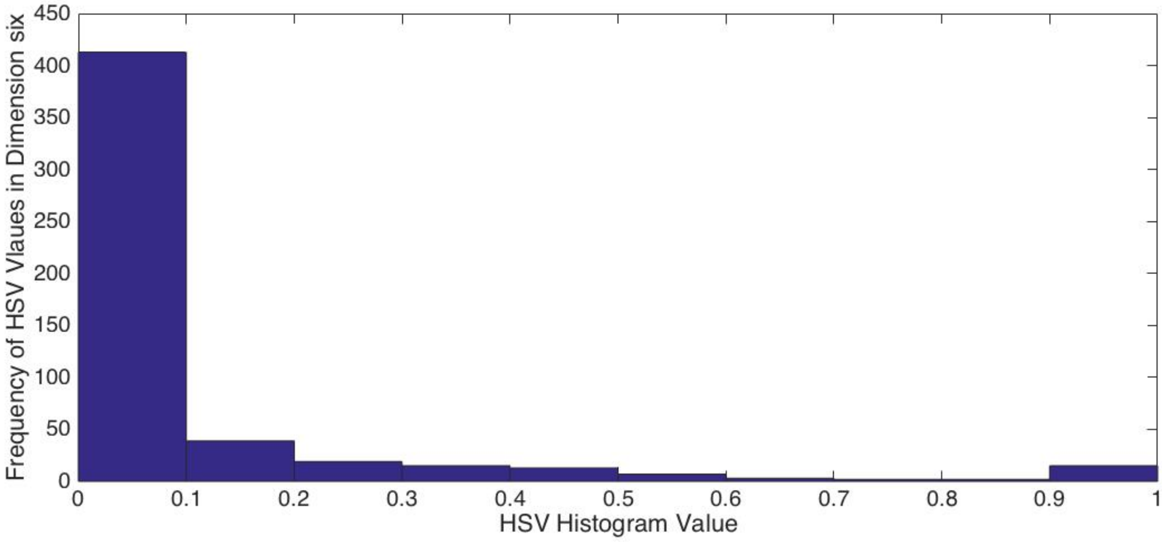


Figure 2. A sample of a skewed distribution from features in Dimension 6 of ebay dataset.

Also similar to $\boldsymbol{\ell_p}$, $\boldsymbol{m_p}$ is under the influence of the dominant dimension when combines the data masses in each dimension for calculating the total dissimilarity. A dimension with very high data mass may influence the total dissimilarity between two instances and result in a situation where other dimensions with lower data masses do not contribute proportionally in the calculating the dissimilarity.

III. PROPOSED NEW DISSIMILARITY

In this section we propose a new dissimilarity to address the limitations with m_p and geometric distance. Our proposed dissimilarity measure will incorporate the effect of region density and geometric distance. The data mass between two instances is used as a proxy for density region. We propose two variants for our new dissimilarity measure as described in the following sections. In the first variant we focus to improve m_p that was proposed in [9] to address the limitations of geometric distance. m_p has its limitations by ignoring the geometric distance, and we propose a weighting using geometric distance to improve the performance of m_p. The second variant focuses on geometric distance and especially ED that is commonly used, as distance measure while it does not consider any impact of data distribution on perceived dissimilarity. To address this limitation, we propose to incorporate the effect of data distribution in the region covering two instances by weighting the geometric distance using the data mass. Also in both variants the weighting helps to moderate the effect of dominant dimension in calculation of total dissimilarity, which was recognised as a limitation with ℓ_p and m_p. In the first variant, the basis of dissimilarity calculation is m_p and geometric distance as a weight to moderate the data mass in certain circumstances, where it would fail to express the actual perceived dissimilarity. In the second variant, geometric distance is the basis for dissimilarity calculation and data mass is used for moderation.

A. Hybrid Data Dependent Dissimialrity

The first variant of the new dissimilarity measure we are proposing is called Hybrid Data Dependent Dissimilarity (HDDD). It uses the geometric distance as a weight for mass-based dissimilarity, where mass-based dissimilarity may not retrieve accurate results. In certain circumstances that mass-based dissimilarity and geometric distance are not consistent, mass-based dissimilarity may fail to retrieve accurate results.

Generally when we calculate the dissimilarity between two instances using m_p four situations may happen.

- **Case 1:** m_p is small due to locating data instances in a sparse region where data mass is low and geometric distance is small too;

- **Case 2-** m_p is small due to locating data instances in a sparse region where data mass is low, but geometric distance is large,
- **Case 3-** m_p is large due to locating data instances in a dense region where data mass is high and geometric distance is large too,
- **Case 4-** m_p is large due to locating data instances in a dense region where data mass is high, but geometric distance is small.

In Cases 1 and 3, m_p and geometric distance are consistent. However, in Cases 2 and 4, their measurements are not consistent and using m_p alone may not be effective. We define the high/low data mass and large/small distance using a threshold, which is the mid-point between minimum and maximum of data mass/distance values between a query and all other data instances in the dataset. Cases 2 and 4 are the situations discussed in the previous section as the limitations of mass-based dissimilarity.

To address the discussed limitations of Cases 2 and 4, we will incorporate the geometric distance between two instances and weight the data mass in each dimension proportionally. So the weight will be applied in two situations. First, in Case 2 where data mass between two instances is low but their distance is large; and second, in Case 4 where data mass between two instances is high but their distance is small.

HDDD is defined as using conventional m_p in Cases 1 and 3, and in Cases 2 and 4, using the weighted m_p as follows:

$$Wm_p(x,y) = \left(\frac{1}{d}\sum_{i=1}^{d} abs\left(W_i \frac{|R_i(x,y)|}{N}\right)^p\right)^{1/p} \quad (3)$$

Where W_i depends on whether l_p and m_p agree in Dimension i. $W_i = 1$ is used in the cases where l_p and m_p agree (i.e, Cases 1 and 3). It is set appropriately to weight the data mass in each dimension proportionally to the geometric distance wherel_p and m_p do not agree. In Case 2 where data mass is low but distance is large, $W_i = abs(x_i - y_i)$ is used to assign a higher weight to the data mass,. In Case 4, where data mass is high between two points but distance is small, $W_i = \frac{abs(x_i - y_i)}{\max_{m\in D} abs(x_i - m_i)}$ is used to assign a lower weight to data mass. D is the set of images in the given database.

In HDDD, we modified m_p and weighted with distance in Cases 2 and 4, and for the rest of situations we used conventional m_p. So basically the calculation of dissimilarity in HDDD is based on m_p. The preliminary results of this method is published in [16].

B. Density Affected Dissimilarity

As discussed in Section II, geometric distance has a limitation by ignoring the data distribution and mass-based dissimilarity may not retrieve the accurate results in certain situations as a result of not considering the geometric distance. In this section, we propose our second variant of our new dissimilarity measure that incorporates geometric distance and data distribution. In this approach, to consider the effect of region density on final perceived dissimilarity, we will moderate the geometric distance between two instances using the density of the region. In the following, we first describe this variant of our new dissimilarity measure, called Desnity Affected Dissimilarity (DAD), and then define the nature of region density effect on final perceived dissimilarity. In the last part, we discuss the proper use of region density as a weight for geometric distance.

1) Density Afftected Dissimilarity

The effect of region density is explained as two instances located in a dense region look more dissimilar compared to locating them in a sparse region. So density of a region can moderate the perceived dissimilarity. Considering this, when dissimilarity between two instances is measured using geometric distance, we need to proportionally weight it based on the density of the region. We propose our new dissimilarity measure as:

$$DAD(x,y) = \left(\sum_{i=1}^{d} \left(abs(x_i - y_i) \times \left(T(|R_i(x,y)|)\right)\right)^p\right)^{1/p} \quad (4)$$

where $T(|R_i(x,y)|)$ is the transformation of data mass between x and y in Dimension i of the feature vector. In the proposed variant, density of the region is used to weight the distance between two instances. If distance is measured in a denser region, then a higher weight will be assigned to that while in sparser region this weight is lower. So the perceived dissimilarity would be different depending on the density of the region that distances is measured.

2) Interaction Effect of Region Density on Geometric Distance

In this section, we explain why the effect of region density on the final perceived dissimilarity is multiplicative. The interaction effect is said to exist when the effect of independent variable on a dependent variable differs depending on the value of a third variable, called moderator variable. This effect is not additive but a multiplicative effect [17-20]. We define the effect of region density on perceived dissimilarity as an interaction effect as follows. In our case, dependent variable is the final perceptual dissimilarity, which depends on the geometric distance between two data points. Geometric distance between two data points is an independent variable and data mass (region density) is the moderator variable. The effect of geometric distance on final perceptual dissimilarity will differ depending on the region density. So an interaction effect exists between geometric distance and region density, which is multiplicative.

Also, multiplication is more robust to outlying values with significantly large/ small ranges than addition, which can be dominated by those outlying values. We can see this varying value ranges in data masses as they can range from the minimum number of two (each defined region between two data instances at least cover those two points) to maximum, the number of data instances in the dataset. In following, we apply this effect as a multiplicative effect in calculation of our proposed variant.

3) Log Transform of Region Density

So far, we have discussed the multiplicative effect of the density of a region on the perceived dissimilarity, and proposed to use it as a weight for geometric distance. So in our proposed dissimilarity, data mass will be multiplied to geometric distance. $|R_i(x, y)|$ in a dataset can be between 2 to N. The minimum data mass is two as the defined region between two instances enclose the two instances and in case of having no data point between them the region include at least two points. So data mass may have very large range dependent on the size of dataset. As shown in equation 4, we consider the effect of region density in each dimension of feature vector, and finally aggregate the dissimilarity in all the dimensions. So in this aggregation process data masses with very large ranges, which represent denser regions, will dominate the final dissimilarity. So the effect of sparse regions/ low data mass, will not contribute proportionally to the final dissimilarity calculation. To consider the region density effectively in our new dissimilarity we have to use a transformation of data mass that balances the over influence of very high and/or very low data mass in some dimensions in the overall dissimilarity.

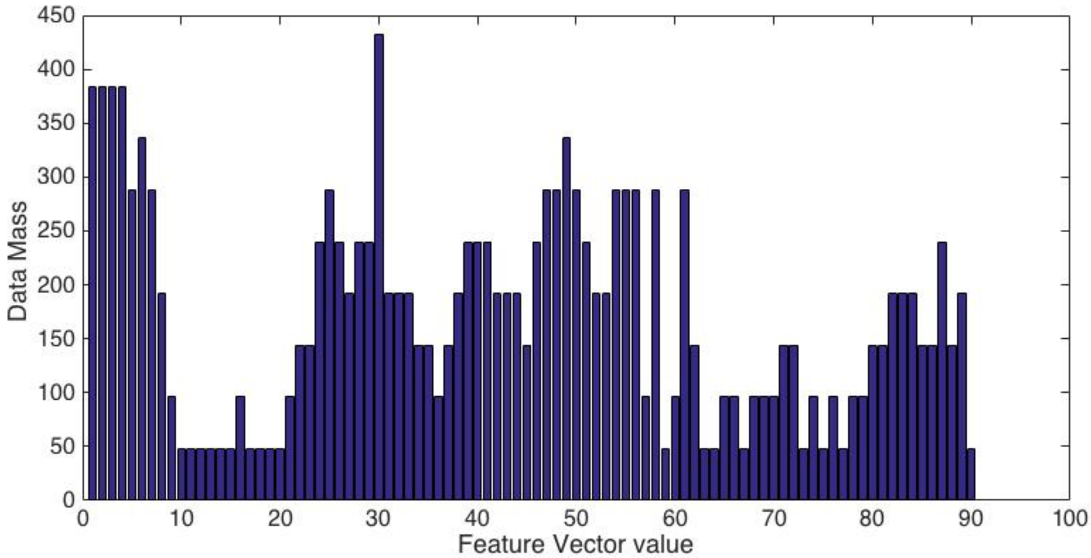


Figure 3. Data masses between two feature vectors.

Logarithmic transform is an established method to deal with highly skewed data distributions [21-23]. The Log transform changes a highly skewed data to a distribution closer to normal and draws out the small numbers. As we mentioned we aim to use data mass as weight for geometric distance and using Log transform balance the contribution of very high and/or very low data mass in some dimensions in the overall dissimilarity. Thus, we used log transformation of $|R_i(x,y)|$, i.e.,$\mathrm{T}(|R_i(x,y)|) = \log|R_i(x,y)|$. Figure 4, shows the Log transform of data masses presented in Figure 3, and we can see not only the distribution is less skewed but also the dimensions with low data masses have a proportional role in the distribution. So Log transform can serve for both of our purposes in rescaling data masses and also give all dimensions a proportional contribution in final dissimilarity estimation as a weight.

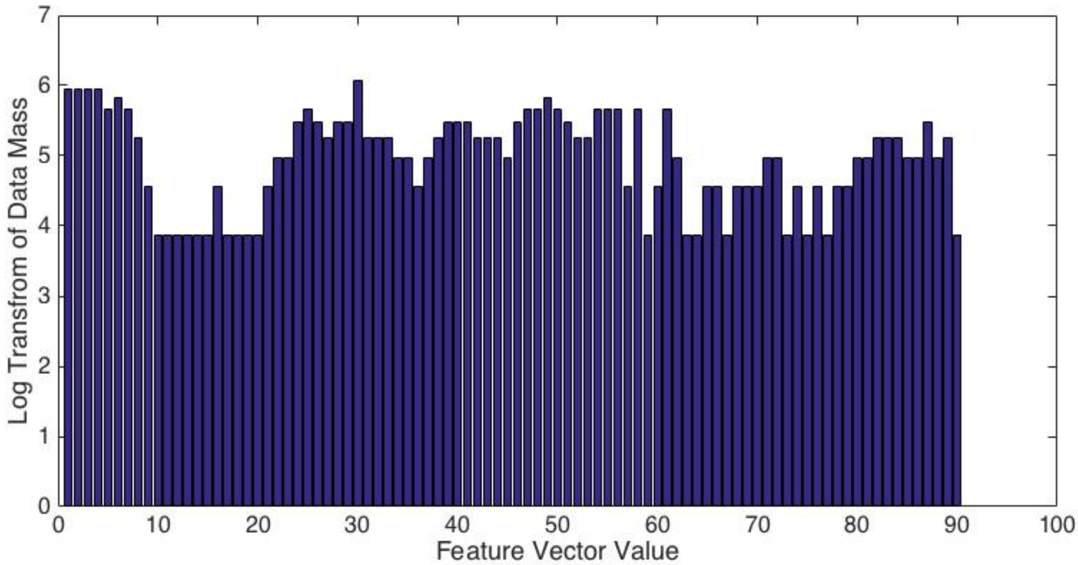


Figure 4. The Log transformed data masses of the ones presented in Figure 3.

IV. FEATURE EXTRACTION

In this section, we describe the features that are used to represent images in this paper. Color is an intuitive and useful feature to represent images. Color histograms are rotation invariant and has been used in image retrieval studies [1, 2]. A colour space is specification of a coordinate system and subspace within that system where a single point represents a distinct colour value. This representation is used for image analysis like extraction of colour histograms. Each colour space has its own merits and demerits depending on the application and hardware specification where it is going to be used. HSV is one of the popular colour spaces. It has three dimensions; hue, which represents the colour, saturation for purity of the colour and value for intensity that ranges from black to white.

As suggested in [24, 25], hue has more importance in distinguishing the perceived colour and two other components has equal importance to represent using colour histogram. So based on literature and discussed HSV colour space, we will give a higher weight to hue and equal weights to saturation and value components. So in this case to apply the weights for HSV colour histogram we will modify Equations 1-4 for calculation of ℓ_p m_p and DAD as follows:

$$l_p(x,y) = \left(\sum_{i=1}^{d} HSV\_W_i \times abs(x_i - y_i)^p\right)^{1/p} \tag{5}$$

$$m_p(x,y) = \left(\frac{1}{d}\sum_{i=1}^{d} HSV\_W_i \times abs\left(\frac{|R_i(x,y)|}{N}\right)^p\right)^{1/p} \tag{6}$$

$$HDDD(x,y) = \left(\sum_{i=1}^{d} HSV\_W_i \times HDDD_i(x,y)^p \right)^{1/p} \tag{7}$$

$$DAD(x,y) = \left(\sum_{i=1}^{d} HSV\_W_i \left((x_i - y_i) \times \left(T\left(Datamass(x_i, y_i)\right) \right) \right)^p \right)^{1/p} \tag{8}$$

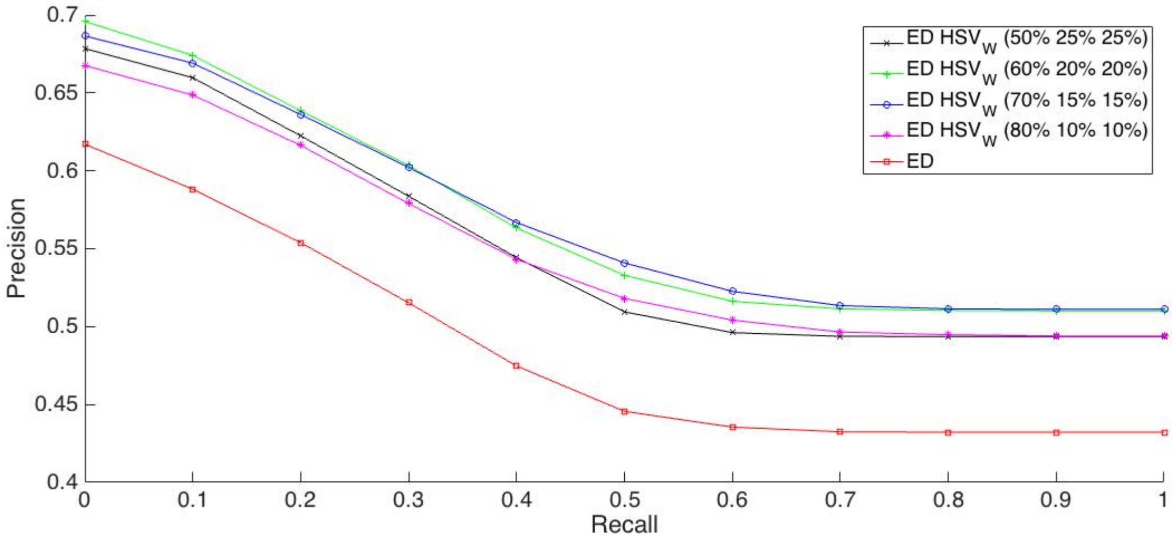


Figure 5. Image retrieval results on ebay dataset using HSV colour historgam with different sets of weights on ED.

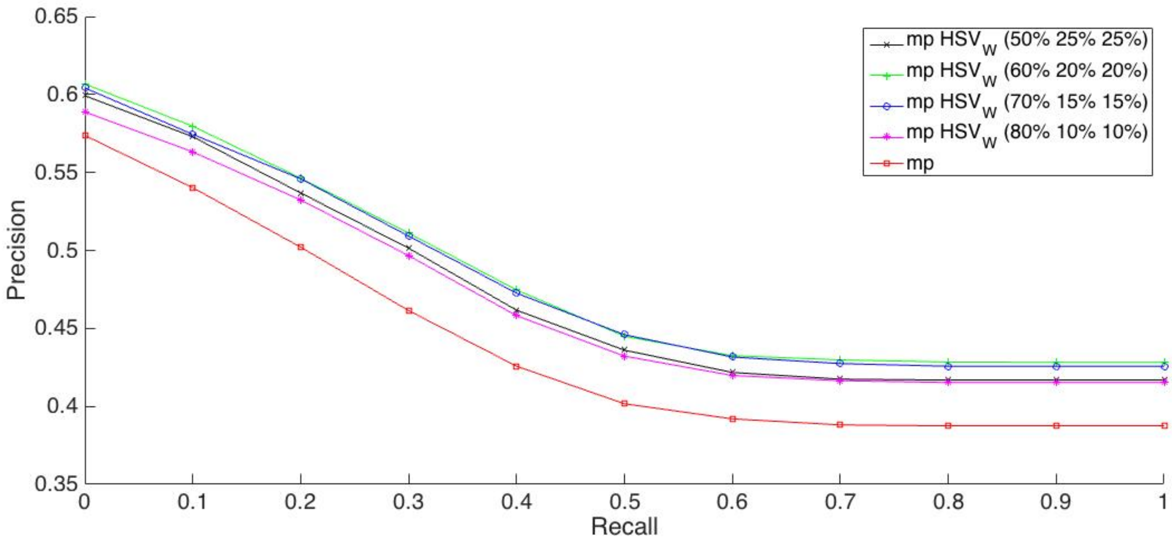


Figure 6. Image retrieval results on ebay dataset using HSV colour historgam with different sets of weights on m_p.

To determine the appropriate weight for different HSV components, we perform image retrieval experiments on ebay [26] dataset using ED and m_p,. We extracted colour histograms of 30 bins for each dimension. Figures 5-6 show the retrieval results and its improvement using different sets of weights. As the weight of 60% for Hue and 20% for Saturation and 20% for Value components showed the optimum performance, we used this setting in this paper

V. EXPERIMENTAL SETUP, RESULTS AND DISCUSSION

In this section, we present and discuss the empirical results on image retrieval. First, we explain our experimental setup and then we compare the performance of the two variants of our proposed dissimilarity, i.e. HDDD and DAD, with m_p and ED.

A. Experimental Setup

In our experiments, images will be presented using HSV colour histogram [24] and we will use the weight (60% H, 20% S, 20% V) as determined in Section IV. The benchmark dataset used in this work is ebay [26] which has the ground truth based on the colour. It has 11 classes. Each class presents one colour and has four categories of objects, 12 images for each object. So it has 48 images in each class and a total of 528 images in the dataset. In this dataset, objects of the targeted colour have been segmented. So our colour histogram will represent only the colour for the main object in the image. Images containing objects with the same colour as that of the query image are considered as relevant images. Figure 8 shows four objects with the same colour from one class and their respective mask images. The colour histogram is extracted from the white section of mask images shown in Figure 7. We will use precision and recall to evaluate the performance of image retrieval in our work. In information retrieval task where an instance can be relevant or non-relevant, precision shows the fraction of retrieved instances that are relevant while recalls is the fraction of relevant instances that are retrieved [27]. The further PR curve is away from the origin, the better is the retrieval performance of the method that curve is representing.

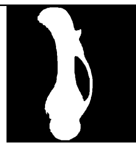
Pink Class	Car	Dress	Pot	Shoe
Original image				
Mask image				

Figure 7. Sample of images and their respected mask image in one class of ebay dataset.

B. Retrieval Results

Figure 8 shows the overall retrieval results using, ED, m_p, HDDD and DAD. Equations 5-8 were used for this experiment in order to consider the HSV colour histogram weights as well. Parameter p was set to 2 in related equations for calculation of m_p, HDDD and DAD. Our experimental results show that DAD performs best overall for image retrieval. ED performs second best, followed by HDDD and then m_p. In the next two sub-section, we will provide a more in-depth comparison and discussion on the performance of HDDD versus m_p, as well as the performance of DAD versus ED. We will also explain why HDDD performs worse than ED.

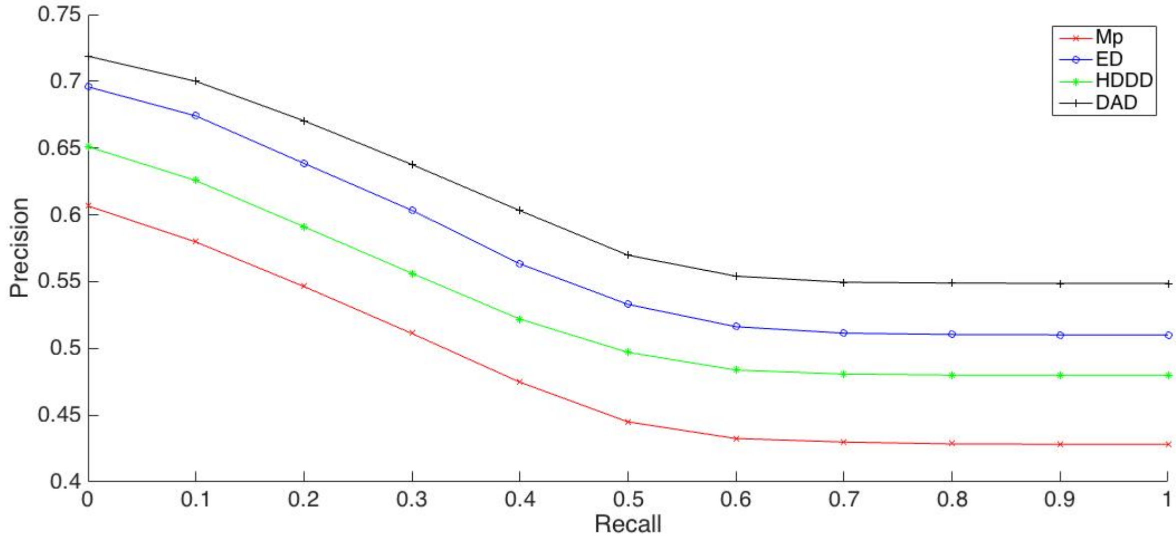
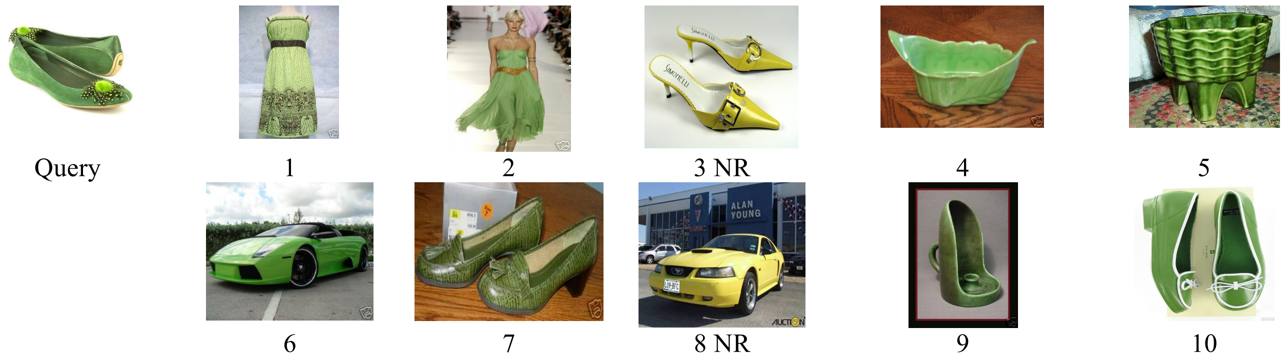


Figure 8. Retrieval results from ebay dataset using different dissimilarity measures.

Figures 9-10 show two visual examples from top 10 retrievals using ED, m_p, HDDD and DAD. Images from the same colour with query are considered as relevant while images with other different colours are considered as non-relevant (NR).

(c) Using $\boldsymbol{m_p}$



(d) Using conventional HDDD

Figure 9. Top 10 retrieval for Query 1

(a) using DAD

(b) Using ED

6 NR 7 8 9 NR 10

(c) Using $\boldsymbol{m_p}$

Query 1 2 3 4 5

6 NR 7 8 NR 9 10

(d) Using conventional HDDD

Figure 10. Top 10 retrieval for Query 1

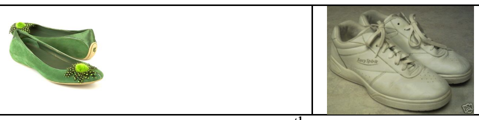
C. Comparison of HDDD and m_p

Figure 8 shows that HDDD preforms better than m_p. This improvement occurred as in HDDD, we incorporate geometric distance with region density where m_p may not to retrieve accurate results. We used weighted m_p as in Equation 3, to address the limitations in Cases 2 and 4. So the basis for dissimilarity calculation in HDDD is m_p and in Cases 2 and 4 it will be weighted using ED.

In this section, to show HDDD obtains better retrieval results than m_p, we show the components that it uses for the calculation of dissimilarity. HDDD uses the data mass between feature vectors of two images and their distance to estimate the final dissimilarity. We show the HDDD dissimilarity in each dimension to illustrate the effect of using distance as the weight in situations where data mass is above the threshold while distance is below that or vice versa. As mentioned before threshold in each dimension is determined as the mid point between the minimum and maximum of data masses/ distances between query and all other data in the dataset.

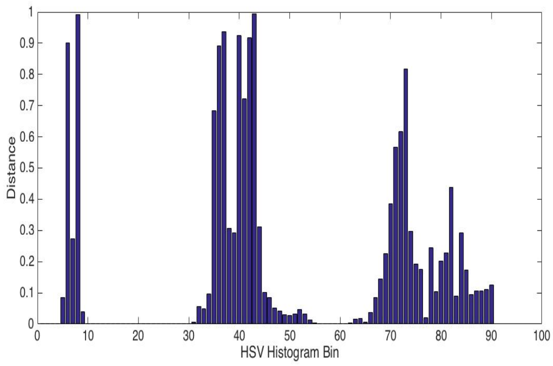
In Figure 9 (c), m_p ranks a relevant image, which appears in the fifth rank of HDDD, lower than 10. It determines that the relevant image: the green pot, is more dissimilar to the query due to the high data mass between them compared to the white shoe which is in its fifth rank. The skewed distribution of features could result in this situation as explained in Section II.B. As an example it is shown in Figures 11 (c-f) that in Dimension 6 of features distance between green shoe and green pot is considerably smaller compared to the white shoe, however the data mass between the green shoe and pot is much higher compared to green and white shoes. So there is an inconsistency between data mass and distance in this dimension that HDDD will adjust that by weighting the data mass accordingly with distance. It is shown in Figure 11 (h) that HDDD weights the data mass by distance in Dimension 6 where the data mass between green shoe and pot is above the threshold while their distance is below the threshold. Using HDDD, the consistency between data mass and distance results in finding the green shoe and pot more similar in dimension six compared to white shoe. Also, HDDD moderates the effect of dominant dimension in calculating the dissimilarity where a dimension with very high data mass could influence the total dissimilarity between green shoe and pot measured by m_p. These can happen in multiple dimensions and finally improve the ranking in HDDD compared to m_p. HDDD has calculated the dissimilarity of 1149.22 between green shoe and pot and ranked it as fifth rank while the white shoe with dissimilarity of 1398.74 has been ranked much lower as 58th.

Figure 12 is another example of inconsistency between data mass and distances in Dimension three of Figures 12 (c, e and g) and Dimension 80 of Figures 12 (d, f and h). The highly skewed data distribution in bin three and 80 is shown in Figures 13-14.

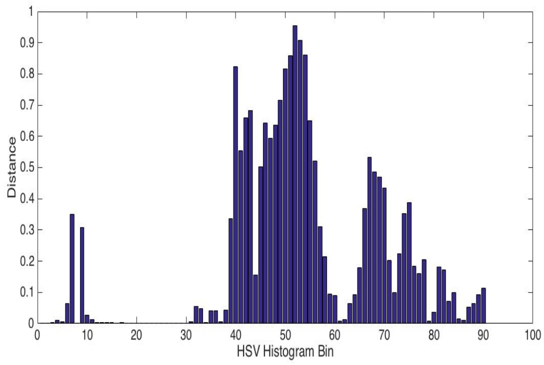


(a) Query (left) and Rank 5th (right) from m_p

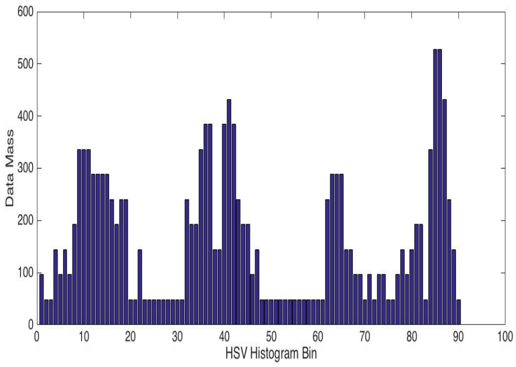
(b) Query (left) and Rank 5th (right) from HDDD



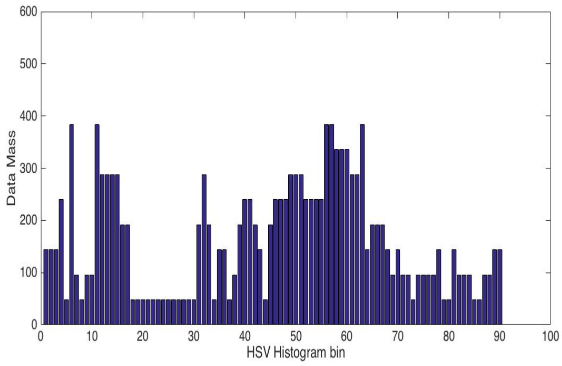
(c) Distance (ED) between images in (a)



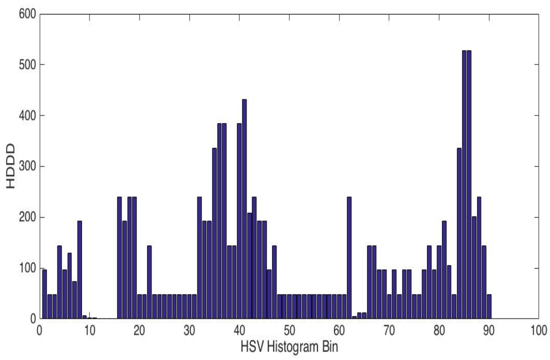
(d) Distance (ED) between images in (b)



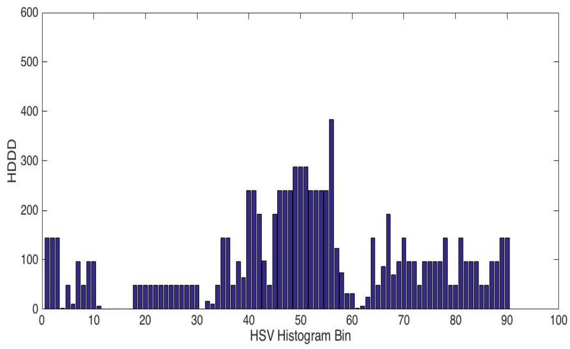
(e) Data mass between images in (a)



(f) Data mass between images in (b)



(g) HDDD between images in (a)

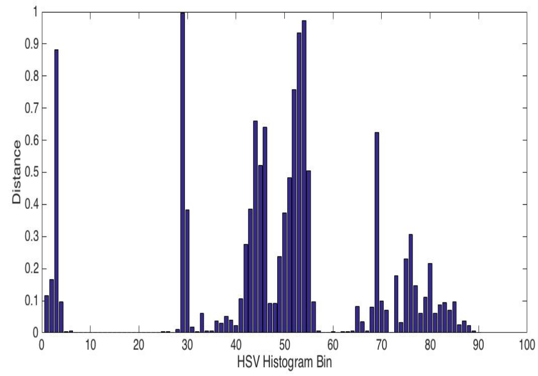


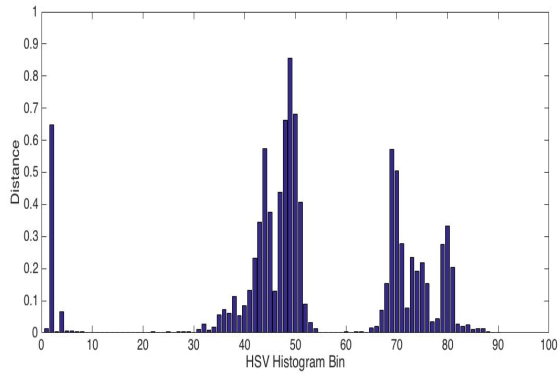
(h) HDDD between images in (b)

Figure 11. HDDD components and final dissimilarity estimation

(a) Query (left) and Rank 1th (right) from m_p

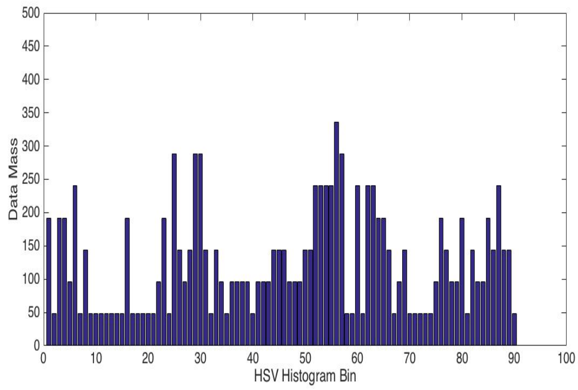
(b) Query (left) and Rank 1th (right) from HDDD

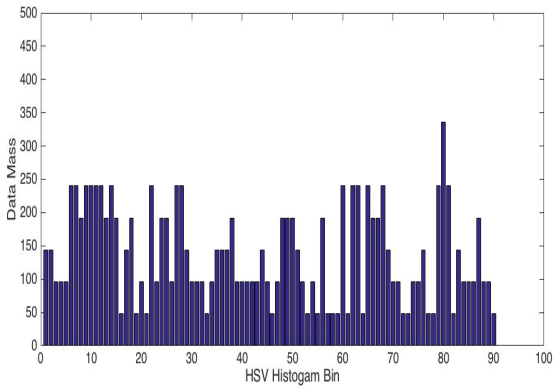




(c) Distance (ED) between images in (a)

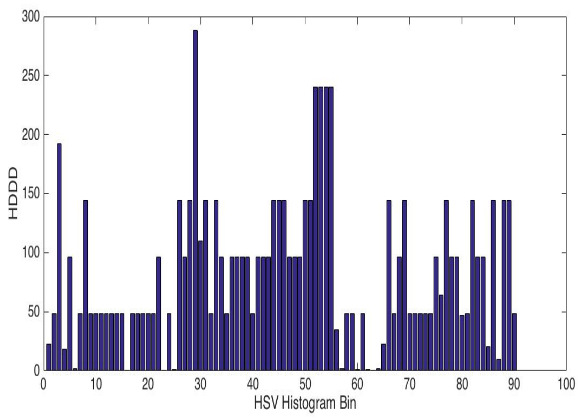
(d) Distance (ED) between images in (b)

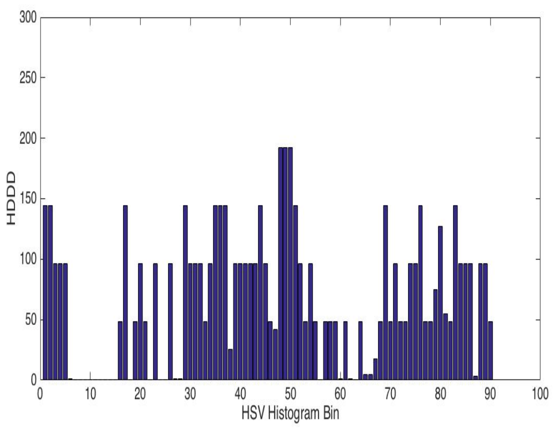




(e) Data mass between images in (a)

(f) Data mass between images in (b)





(g) HDDD between images in (a)

(h) HDDD between images in (b)

Figure 12. HDDD components and final dissimilarity estimation

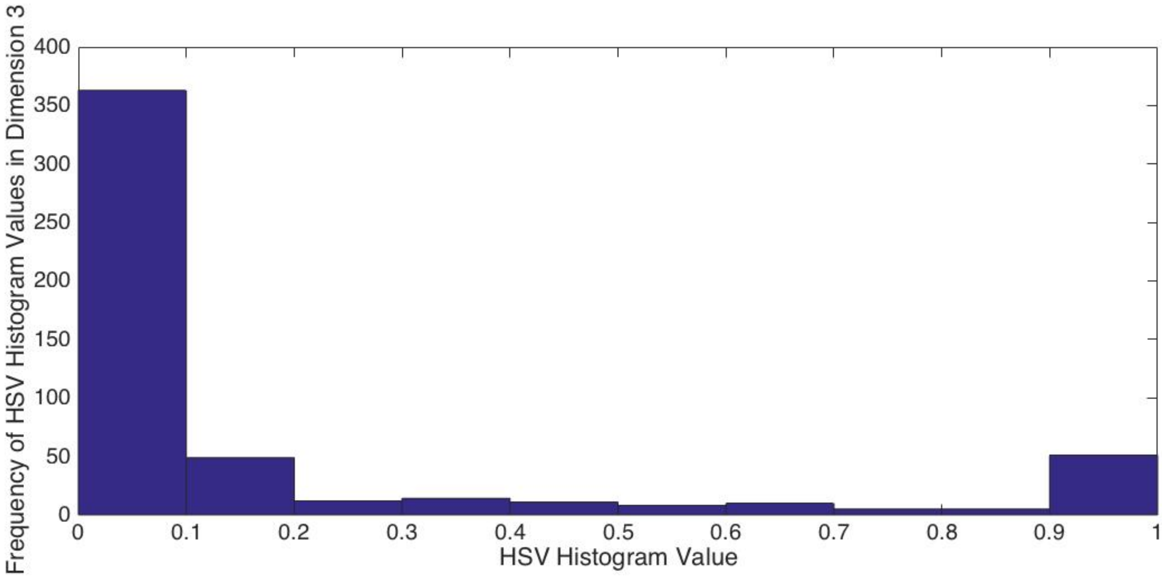


Figure 13. Distribution of features in Dimension 3 of dataset.

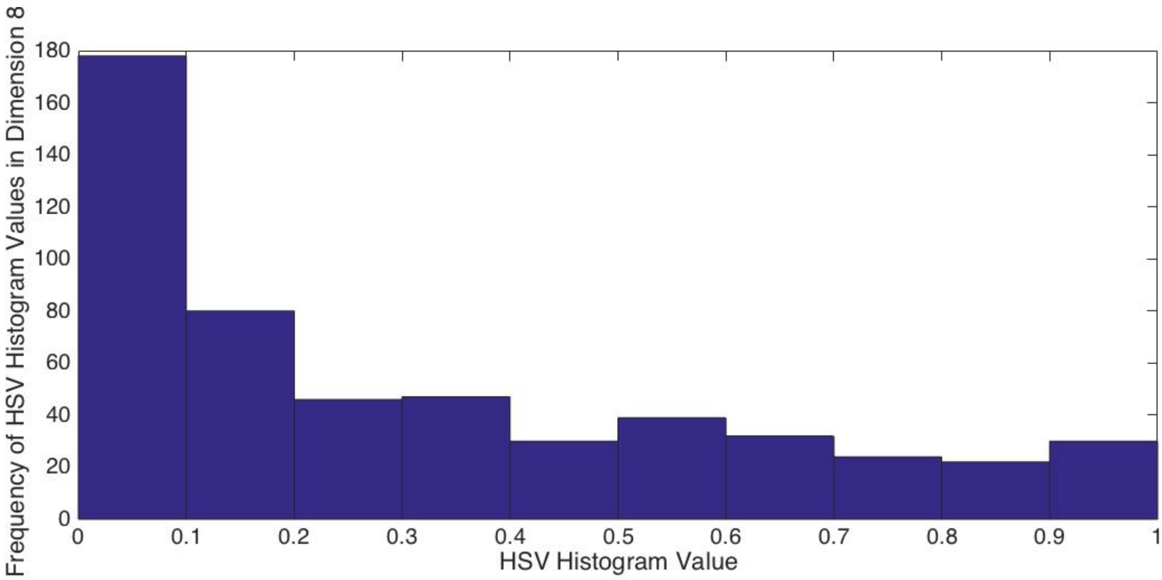


Figure 14. Distribution of features in Dimension 80 of dataset.

HDDD has not performed better than ED. It occurred as weighted m_p that incorporates distance and region density is applied where data mass and distance are not consistent in Cases 2 and 4. In the discussed cases data mass between two instances is high/low while the respected distance is small/large. To determine the high/ low data mass and small/ large distance a threshold is defined which is the mid point between minimum and maximum of data masses/ distances between a query point and all data points in the dataset. The defined threshold may raise a potential limitation as the point just below and above the mid point will be considered as low and high data mass or small and large distance. However these two border points are very close and their difference does not represent the actual difference between low and high data mass or small and large distance. It considers a border point just below the threshold as located in a low data mass area (sparse region) and the one just above the threshold as located in a dense area however they have very similar data masses. The same situation is for distance values.

D. Comparison of DAD and ED

In this section we discuss the performance of ED and DAD, through visual examples in Figures 9-10. To discuss the performance of DAD, we show the components that it uses for calculating dissimilarity between two instances, which are distance and Log transformation of data mass. The basis for calculation of DAD is geometric distance, which is moderated by region density as the weight. We show DAD in each dimension, to illustrate how region density could moderate the geometric distance between two images in the feature space.

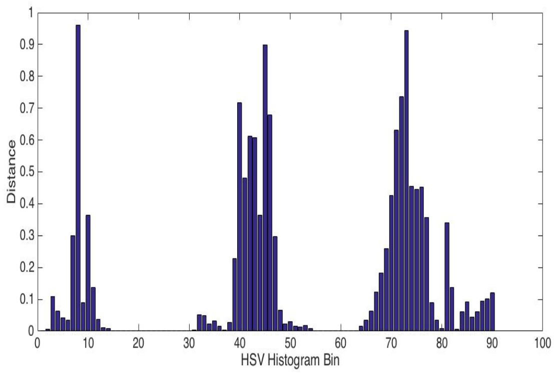
As shown in Figures 9-10, DAD as the dissimilarity measure could retrieve all images from the same class with query compared to ED and m_p and HDDD. For example in Figure 9 (a) using DAD all the top 10 retrieved images are from the class of green, the same colour with the query, however in Figure 9 (b-d), ED and m_p and HDDD retrieved images from other colours such as blue, pink, yellow. Figure 10 (a) also shows that top 10 retrieval using DAD are from the same class with query which is brown colour, compared to Figures 10 (b-d) that have more retrievals from other colours such as, grey red, purple and pink.

DAD improved retrieval results by considering both geometric distance and the effect of region density in estimating the final dissimilarity between two images. Here we use the visual examples to compare the performance DAD and ED. We choose ED as it is the basis for dissimilarity calculation in DAD and it showed the second best performance after DAD in our retrieval results.

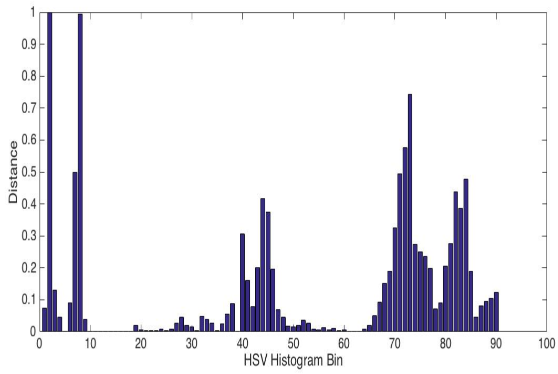
In Figure 9 (b), ED has ranked the green pot as a relevant image in its seventh rank lower than a pink pot as a non-relevant image. This occurred as ED relies only on the distance between colour histogram of query and these two images and does not consider the data distribution. The green pot has been ranked higher in retrievals from DAD in Figure 9 (a) and pink pot ranked lower than 10. Unlike ED, DAD considers if two instances are located in a dense/ sparse region their distance will be perceived differently.


(a) Query (left) and Rank 5th (right) from DAD

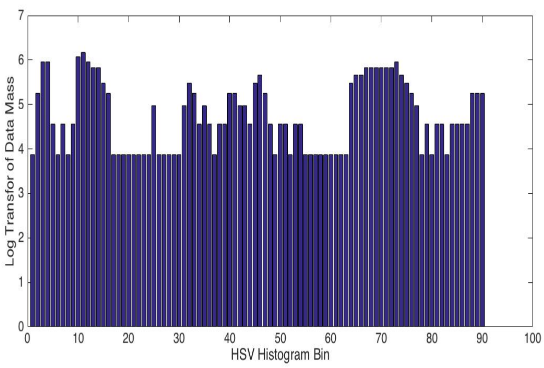

(b) Query (left) and Rank 6th (right) from ED



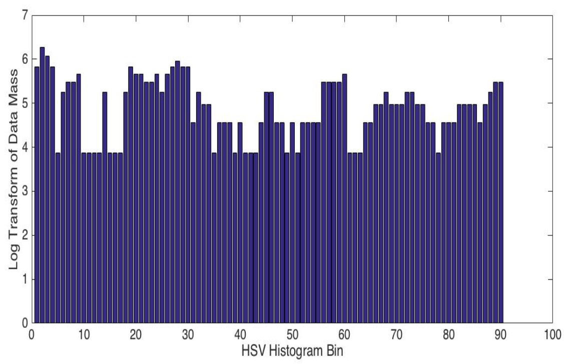
(c) Distance (ED) between images in (a)



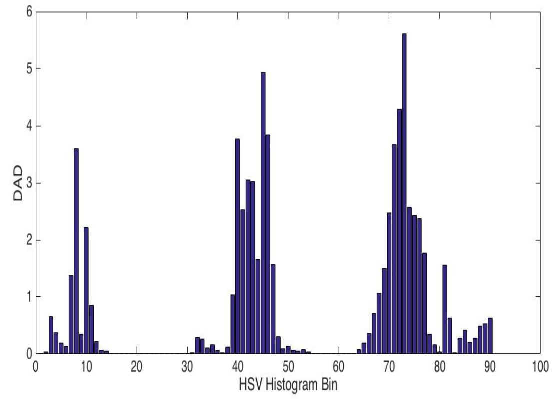
(d) Distance (ED) between images in (b)



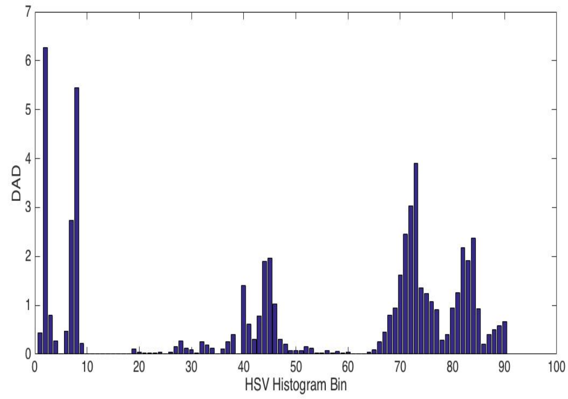
(e) Log transfrom of data masses between images in (a)



(f) Log transfrom of data mass between images in (b)



(g) DAD between images in (a)



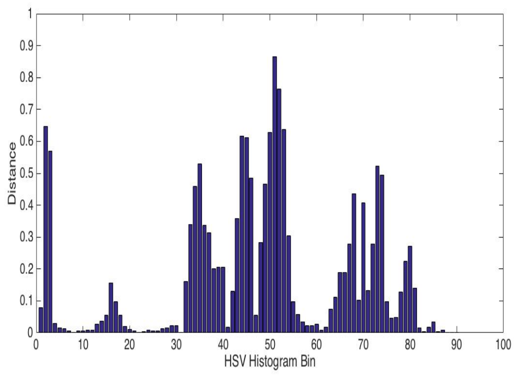
(h) DAD f between images in (b)

Figure 15. Comparison of ED and data mass between two feature vectors and their effect on DAD.

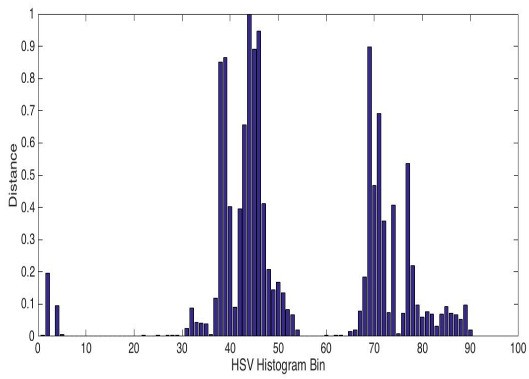
Here we use an example to show how DAD is calculated and improved the results. Figures 15 (c-d) show that the distance between the query and the green and pink pots in Dimension 8 are very similar, 0.96 and 0.99. However in this dimension the distance between the query and green pot has been measured where two points have been located in a denser area (higher data mass) as it is shown in Figure 15 (f), compared to the same bins in Figure 15 (e) which is located in a sparser region. The Log of data mass between green shoe and green pot in Dimension 8 is 3.8 while it is 5.4 for the pink pot. Using the Log transformation of data mass as the proxy for region density gave higher weight to the distance, which is measured in a denser area and vice versa. DAD in Figures 15 (g-h) shows the weighted distances, which means geometric distance is moderated by the region density. The weighted distances in Dimension 8 between the green shoe, green and pink pots are 3.6 and 5.34. Although they have similar distances, using the density of the region to moderate the distance resulted that pink pot in Dimension 8 be found much more dissimilar to the green shoe compared to the green pot. Also, the weighting in DAD helped to moderate the effect of dominant dimension in calculating the distance where in Dimension 8 the large distance between green shoe and pot could influence the total distance measured by ED. Considering this weighting in other dimensions that moderate the distances resulted in the dissimilarity of 7.3 between the green shoe and pot while the dissimilarity of 7.8 for the pink pot ranked it lower than 10 in retrievals from DAD. Figure 16 is another example that illustrates the same scenario.

(a) Query (left) and Rank 7th (right) from DAD

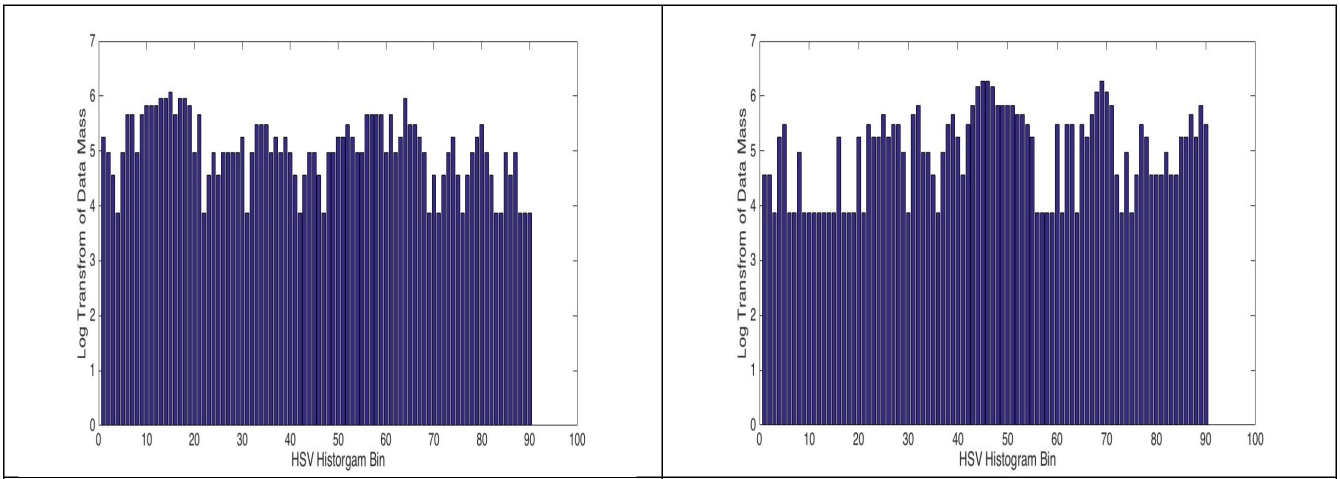
(b) Query (left) and Rank 7th (right) from ED



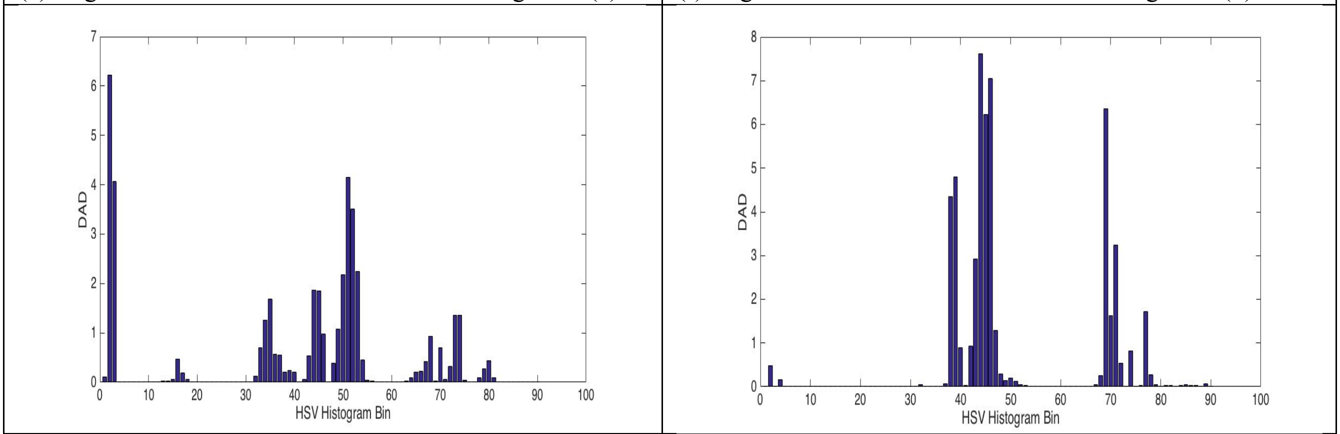
(c) Distance (ED) between images in (a)



(d) Distance (ED) between images in (b)



(e) Log transfrom of data masses between images in (a)	(f) Log transfrom of data mass between images in (b)



(g) DAD between images in (a)	(h) DAD between images in (b)

Figure 16. Comparison of ED and data mass between two feature vectors and their effect on DAD.

VI. CONCLUSION

In this work, we studied the strengths and limitations of geometric distance and mass-based dissimilarity. Geometric distance relies only spatial distance of two instances and ignores the data density/ distribution in the dataset. However, density of the region has an effect on the perceived dissimilarity. On the other hand mass-based dissimilarity only considers the data distribution/ region density and completely ignores the geometric distance that measures the dissimilarity to a certain extent. Also both geometric distance and mass-based dissimilarity are under the influence of dominant dimension when they combine the dissimilarities in each dimension. We used the strengths of both measures and incorporated the geometric distance and region density into a new dissimilarity measure. We defined two variants for the proposed dissimilarity, HDDD and DAD. In the first one, the basis for calculation of dissimilarity is m_p, which is weighted by geometric distance in Cases 2 and 4 where it may fail to express their actual perceived dissimilarity. In the second variant, geometric distance is basis for calculation of dissimilarity and region density as the weight is used to moderate it in the final perceived dissimilarity.

The new hybrid dissimilarity measure considers the effect of region density on geometric distance and the final perceived dissimilarity. The weighting used in the new hybrid dissimilarity also moderates the effect of dominant dimension in dissimilarities measured by ℓ_p and m_p. We evaluated our proposed dissimilarity measure in image retrieval and it could perform better than m_p and ED. The proposed dissimilarity has the potential to be used in other applications such as information retrieval and classification.

REFERENCES

1. Zhang, D., M.M. Islam, and G. Lu, *A review on automatic image annotation techniques.* Pattern Recognition, 2012. **45**(1): p. 346-362.
2. Zhang, D. and G. Lu, *Review of shape representation and description techniques.* Pattern Recognition, 2004. **37**(1): p. 1-19.
3. Teng, S.W., M.T. Hossain, and G. Lu, *Multimodal image registration technique based on improved local feature descriptors.* Journal of Electronic Imaging, 2015. **24**(1): p. 013013.
4. Hossain, M.T., et al. *An enhancement to SIFT-based techniques for image registration*. in *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*. 2010. IEEE.
5. Aryal, S., et al., *Data-dependent dissimilarity measure: an effective alternative to geometric distance measures.* Knowledge and Information Systems, 2017: p. 1-28.
6. Bosch, A., A. Zisserman, and X. Muñoz, *Scene classification via pLSA*, in *Computer Vision–ECCV 2006*. 2006, Springer. p. 517-530.
7. Jain, A.K. and A. Vailaya, *Image retrieval using color and shape.* Pattern Recognition, 1996. **29**(8): p. 1233-1244.
8. Kim, H.-K., et al. *A modified Zernike moment shape descriptor invariant to translation, rotation and scale for similarity-based image retrieval*. in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. IEEE.
9. Aryal, S., et al. *Mp-Dissimilarity: A Data Dependent Dissimilarity Measure*. in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE.
10. Krumhansl, C.L., *Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density.* 1978.
11. Rogers, P.L., *Encyclopedia of distance learning*. 2009: IGI Global.
12. Zhang, D. and G. Lu. *Evaluation of similarity measurement for image retrieval*. in *Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on*. 2003. IEEE.
13. Malik, F. and B. Baharudin, *Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain.* Journal of King Saud University-Computer and Information Sciences, 2013. **25**(2): p. 207-218.
14. Khotanzad, A. and Y.H. Hong, *Invariant image recognition by Zernike moments.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1990. **12**(5): p. 489-497.
15. Singh, C., *Local and global features based image retrieval system using orthogonal radial moments.* Optics and Lasers in Engineering, 2012. **50**(5): p. 655-667.
16. H.Shojanazeri, S.W.T., Dengsheng Zhang, Guojun Lu, *A Hybrid Data Dependent Dissimilarity Measure for Image Retrieval*, in *Digital Image Computing: Techniques and Applications (DICTA)*. 2017, IEEE: Sydney, Australia.
17. Bailey, R.A., *Design of comparative experiments*. Vol. 25. 2008: Cambridge University Press.
18. Cox, D.R., *Planning of Experiments, New York: JohnWiley.* CoxPlanning of Experiments1958, 1958.
19. Jaccard, J. and R. Turrisi, *Interaction effects in multiple regression*. 2003: Sage.
20. Foulger, D. *Credibility and the statistical interaction variable: Speaking up for multiplication as a source of understanding*. in *Annual Conference of the Speech Communication Association, San Antonio, TX. Retrieved from* http://davis. *foulger. info/papers/statisticalInteraction1979. htm*. 1979.
21. Osborne, J.W., *Improving your data transformations: Applying the Box-Cox transformation.* Practical Assessment, Research & Evaluation, 2010. **15**(12): p. 1-9.
22. Warton, D.I. and F.K. Hui, *The arcsine is asinine: the analysis of proportions in ecology.* Ecology, 2011. **92**(1): p. 3-10.
23. Keene, O.N., *The log transformation is special.* Statistics in medicine, 1995. **14**(8): p. 811-819.
24. Sural, S., A. Vadivel, and A.K. Majumdar, *Histogram generation from the HSV color space*, in *Encyclopedia of Information Science and Technology, First Edition*. 2005, IGI Global. p. 1333-1337.
25. Zhao, Q., et al. *Stone images retrieval based on color histogram*. in *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*. 2009. IEEE.
26. Van De Weijer, J., C. Schmid, and J. Verbeek. *Learning color names from real-world images*. in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. 2007. IEEE.
27. Buckland, M. and F. Gey, *The relationship between recall and precision.* Journal of the American society for information science, 1994. **45**(1): p. 12.